ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

«ΔΙΟΙΚΗΣΗ, ΑΝΑΛΥΤΙΚΗ ΚΑΙ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ ΕΠΙΧΕΙΡΗΣΕΩΝ»

Master of Science in

Business Administration, Analytics and Information Systems

Master Thesis

Machine Learning and classification of transport mode choice using Python

Marios Melachroinos

Supervisor

Thanasis Argyriou

Athens 2024

The approval of this thesis by the Department of Economics (MSc Administration, Analytics and Information Systems) of the National and Kapodistrian University of Athens does not necessarily imply the acceptance of the author's views on behalf of the Department.

I hereby certify the submitted thesis and the work presented is personal and that all sources and materials used have been properly referenced in the text and bibliography.

Marios Melachroinos

Acknowledgements

I would like to express my heartfelt gratitude to my family for their unwavering support and encouragement through this journey. Their belief in my abilities and constant encouragement provided the motivation needed to navigate the challenges of this master's thesis.

I extend my sincere appreciation to Professor Thanasis Argyriou for his invaluable guidance and mentorship throughout this research endeavour. His expertise, constructive feedback, and commitment to academic excellence have been instrumental in shaping the quality of this thesis.

I am also thankful to my friends and colleagues who offered their insights and assistance during various stages of this research. Their collaboration added depth to the project and enhanced the overall learning experience.

This thesis is a testament to the collective efforts of those who have supported and inspired me. Their contributions have played a pivotal role in the successful completion of this academic pursuit.

CONTENTS

TABLE OF FIGURES

LIST OF TABLES

ABSTRACT

Urban transportation systems play a pivotal role in shaping modern cities, influencing economic productivity and environmental sustainability. As cities grapple with expanding populations and evolving transportation options, understanding the dynamics behind individuals' mode choices becomes crucial. This thesis navigates the intersection of urban mobility and data science, employing machine learning and Python programming to classify transport mode choices. While traditional methods rely on statistical analyses, machine learning offers a paradigm shift by enhancing accuracy through vast dataset analysis.

Focusing on the city of Thessaloniki, this thesis explores machine learning and classification of mode choice in the context of residents' commuting behaviour. The data were collected through a survey both delivered online and in paper form. After cleaning and preprocessing the data, four models were applied to predict commute mode. Decision Tree, Random Forest, XGBoost and Stacked model. Different iterations were used to find optimal performance, such as feature reduction and parameter tuning. The best results were obtained by XGBoost model, followed by Random Forest. Distance of the trip and time required for commuting to work were the most important features for the models. For evaluating extended features and their importance more data are required for training the models.

ΠΕΡΙΛΗΨΗ

Τα αστικά συστήματα μεταφοράς διαδραματίζουν ένα κεντρικό ρόλο στον σχεδιασμό των σύγχρονων πόλεων, επηρεάζοντας την οικονομική παραγωγικότητα και την περιβαλλοντική βιωσιμότητα. Καθώς οι πόλεις αντιμετωπίζουν την αύξηση του πληθυσμού και την εξέλιξη των επιλογών μεταφοράς, καθίσταται ζωτικής σημασίας η κατανόηση των δυναμικών που καθοδηγούν τις ατομικές επιλογές μεταφοράς. Αυτή η διατριβή επικεντρώνεται στον συνδυασμό της αστικής κινητικότητας και της επιστήμης δεδομένων, χρησιμοποιώντας μηχανική μάθηση και τη γλώσσα προγραμματισμού Python για την ταξινόμηση των επιλογών μεταφοράς. Ενώ οι παραδοσιακές μέθοδοι βασίζονται σε στατιστικές αναλύσεις, η μηχανική μάθηση προσφέρει έναν παράδειγμα μεταστροφής παραδειγματικότητας μέσω της ανάλυσης μεγάλων συνόλων δεδομένων.

Επικεντρωμένη στην πόλη της Θεσσαλονίκης, αυτή η διατριβή εξερευνά τη μηχανική μάθηση και την κατηγοριοποίηση της επιλογής τρόπου μετακίνησης στο πλαίσιο της καθημερινής συμπεριφοράς των κατοίκων. Τα δεδομένα συλλέχθηκαν μέσω μιας έρευνας που διανεμήθηκε τόσο online όσο και σε έντυπη μορφή. Μετά τον καθαρισμό και την προ επεξεργασία των δεδομένων, τέσσερα μοντέλα εφαρμόστηκαν για την πρόβλεψη του τρόπου μετακίνησης. Δέντρο απόφασης, Τυχαίο Δάσος, XGBoost και Ένωση μοντέλων. Χρησιμοποιήθηκαν διάφορες επαναλήψεις για την εύρεση της βέλτιστης απόδοσης, όπως η μείωση χαρακτηριστικών και η ρύθμιση παραμέτρων. Τα καλύτερα αποτελέσματα προήλθαν από το μοντέλο XGBoost, ακολουθούμενο από το Τυχαίο Δάσος. Η απόσταση του ταξιδιού και ο χρόνος που απαιτείται για την μετακίνηση στην εργασία ήταν τα πιο σημαντικά χαρακτηριστικά για τα μοντέλα. Για την αξιολόγηση παραπάνω χαρακτηριστικών και τη σημασία τους απαιτούνται περισσότερα δεδομένα για την εκπαίδευση των μοντέλων.

1. INTRODUCTION

Urban transportation systems are the backbone of modern cities, influencing everything from economic productivity to environmental sustainability. As the world grapples with the challenges of expanding urban populations and evolving transportation options, understanding the dynamics behind individuals' choices of transportation modes becomes increasingly crucial. This thesis embarks on a journey into the intersection of urban mobility and data science, leveraging machine learning techniques and the programming language Python to classify transport mode choices. In a field of a rapidly changing urban landscape, transportation modes have diversified beyond conventional choices like cars and public transit to encompass an array of options, including ridesharing and micro-mobility solutions. Traditional methods of understanding and predicting transportation behaviours have relied on statistical analyses, but the emergence of machine learning offers a paradigm shift. The capacity to analyse vast datasets and different patterns provides a new opportunity to enhance the accuracy and efficiency of predicting transport mode choices. However, even with the enhanced accuracy achieved by machine learning models in predicting transport mode choices, their applicability is often limited to specific cases. A model tailored to a particular transport behaviour may not generalize well to other scenarios. This limitation arises due to substantial variations in infrastructure, city layouts, and resident preferences across different regions. To gain a comprehensive understanding of transport mode behaviour and patterns, it is essential to apply models to diverse cases and scenarios.

This Thesis delves into the realm of machine learning and classification of mode choice with a specific focus on how residents in the city of Thessaloniki commute to work. Despite being the second-largest city in Greece and hosting major transportation hubs like a significant port and an airport, Thessaloniki's infrastructure is often characterized by mediocrity, offering only buses as the available option for public transportation. The primary goal of this thesis is to utilize machine learning techniques, leveraging Python programming language, to predict commuting behaviour in Thessaloniki. Python is a high-level programming language known for its simplicity and readability. The main strength of the language is the extensive pool of libraries and packages that constitute machine learning techniques and other developing tasks easier to implement (python.org, 2024).

The subsequent chapters will unfold a comprehensive exploration of existing literature, laying the groundwork for understanding the historical context and evolution of transportation mode choice research. Additionally, the terminology, models, and evaluation metrics in machine learning will be explored upon, drawing insights from literature papers and books. The methodology section will detail the approach taken in data collection process, model selection and evaluation techniques, followed by a comprehensive exploratory data analysis. Subsequent chapters will concentrate on the practical application of models and the interpretation of their outputs. Ultimately, conclusive remarks will be

drawn regarding the results and limitations of the study, accompanied by recommendations for future research and endeavours.

2. LITERATURE REVIEW

Forecasting an individual's choice of transportation mode has garnered significant academic interest in recent years. The field of transportation and behavioural analysis has primarily relied on the extensive utilisation of logit models in existing research. In general, logit models and logistic regression analysis have been employed, specifically to examine the connection between the likelihood of binary or ordinal responses and explanatory variables using the maximum likelihood estimation method (Trueck, 2009). According to Trueck (2009), the logistic function is given by the following expression:

$$P(Y = 1 \mid x_1, \ldots, x_k) = \frac{exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}{1 + exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}.$$

The above can also be rewritten as:

$$logit(P(Y = 1 \mid x_1, \ldots, x_k)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

McFadden (1972, 1974) posed a significant challenge to the initial approach to comprehending travel behaviour. In his exploration of the analysis of transit behaviour, he advocated for the application of the multinomial logit model. Like binary logit models, the multinomial model adheres to the same principles and assumptions, with the notable difference being the presence of multiple alternative choices (Lee & Kim, 2023). Furthermore, because of the assumption of independence among the options, the total probabilities of all choices add up to 1 (Lee & Kim, 2023). Following the introduction of the logit model, subsequent research endeavours have sought to extend this model and investigate travel behaviour, thereby addressing certain constraints associated with the original logit model. Such examples are the nested logit model (Willis, 2014) and the mixed logit model proposed by McFadden & Train (2000). The nested logit model organises choices into separate nests and permits varying correlations between these nests. As a result, correlations are consistent within each nest, but for options situated in different nests, the unobservable elements are uncorrelated and, in fact, entirely independent (Willis, 2014). The mixed logit model is defined as a standard multinomial model, also called “latent class model”, where the coefficients are chosen from a cumulative distribution, introducing an element of randomness (McFadden & Train, 2000).

2.1. Factors influencing transport mode selection.

Since the appearance of logit models, numerous studies have surfaced with the objective of delving into the crucial factors that affect transportation mode selection, going beyond the conventional determinants like cost or travel distance. In a study conducted by Mayo and Taboada in 2020, they employed a hierarchy model to assess the factors influencing the choice of public transportation mode among respondents in the Philippines. Their survey results indicated that safety was the most significant

consideration, followed by cost, comfort, and concerns about environmental sustainability (Mayo & Taboada, 2020). Another study by Donkor et al. in 2020 examined the role of emotions in transport mode selection, focusing on respondents in the city of Edinburgh. Their findings revealed that an individual's feelings and experiences related to public transportation, along with their socio-demographic characteristics, exerted a substantial influence on their transit behaviour (Donkor et al., 2020). Additionally, McCarthy et al. in 2017 conducted further research suggesting that the presence of young children in a family had an impact on transportation behaviour. Specifically, they proposed that families with children preferred car usage over other sustainable transit options, with psychosocial factors and household characteristics playing pivotal roles in the choice of transportation mode (McCarthy et al., 2017). The influence of various weather conditions in transport mode choice have also been explored in the literature. Bocker et al. (2016) delved into weather-related factors and their connection to transit choices. Based on their analysis of travel diaries in the Netherlands, their results indicate that individuals who opt for walking or cycling modes are particularly affected by weather conditions (Bocker et al., 2016).

Additional research in the field of transportation focuses on mode selection during the COVID-19 pandemic and its impact on individual transit choices. To elaborate, Mussone & Changizi (2023) conducted a study that utilised a multinomial logistic regression model to investigate the factors influencing transportation mode choices prior, during, and post COVID-19 lockdowns. Their research, based on data from residents in Milan, Italy, indicated that socio-demographic factors, as well as individual preferences and concerns related to public transportation, played the most significant roles in predicting transport mode choices during the pandemic (Mussone & Changizi, 2023). It is worth noting that, despite mandatory contamination control measures, many residents in various countries expressed heightened concerns about the spread of the virus within public transportation during and after lockdown restrictions. Those concerns led residents to shift from relying on public transportation to utilising private vehicles for their commuting requirements. The phenomenon is further investigated by Das et al. (2021), who employed a logistic regression model to examine travel behaviour and modal transitions. Their research findings indicate that demographic factors exert a considerable influence on preferences for switching transportation modes. Additionally, trip-related factors, including travel time and health conditions, demonstrate a robust association with the inclination to shift from public transportation to using cars (Das et al., 2021).

2.2. Machine Learning essentials

An alternative to the traditional logit models, comes with the introduction of machine learning. According to Zhou (2021), Machine learning is a method that enhances the performance of systems through computational learning from prior experiences. In the realm of computer systems, these experiences are embodied in the form of data. The central objective of machine learning is to create

learning algorithms capable of constructing models based on this data. When the learning algorithm is supplied with experiential data, it yields a model capable of making predictions for new observations (Zhou, 2021). Based on the presence or absence of labelled training data, learning problems can be categorised into two groups: supervised and unsupervised learning. Supervised learning encompasses a training phase in which the algorithm is supplied with a dataset comprising pairs of input and corresponding output (referred to as labelled data). During this phase, the algorithm acquires the ability to make predictions or classifications by drawing insights from this labelled data (Zhou, 2021). In supervised learning, the primary tasks are categorised into regression and classification, depending on whether the prediction output is continuous or discrete. In the case of classification problems, when there are only two possible labels, it is referred to as a "binary classification problem" (Zhou, 2021). If there are multiple possible labels, it is termed a "multi classification problem" (Zhou, 2021). Common algorithms that aim to solve regression and classification problems include Naïve Bayes Classifier, K-Nearest Neighbours, Decision Trees, Ensemble Learning, Boosting, Support Vector Machines and Neural Networks (Kubat, 2021).

In contrast, unsupervised learning is a category of machine learning in which the algorithm is given input data but does not have access to predefined output labels. In this context, the algorithm's role is to autonomously identify patterns, structures, or relationships within the data, all without prior knowledge of the expected output (Zhou, 2021). The main task in unsupervised learning is clustering which involves the process of categorising data points by identifying their similarities (Zhou, 2021). The most common technique for such problems is K-Means Clustering.

There is also a third form of learning called “Reinforcement learning” (Kubat, 2021). In this domain, the objective differs significantly. Here, the agent's role is not to induce knowledge from a pre-classified dataset but to engage in active experimentation with a system. The system, in turn, provides feedback in the form of rewards or penalties in response to the agent's actions. The agent's primary aim is to refine its behaviour by seeking to maximise rewards and minimise penalties as it interacts with the system (Kubat, 2021).

2.3. Classification Algorithms & Techniques

The most common algorithms and machine learning techniques that aim to solve classification tasks in a supervised problem are presented as followed:

K-Nearest Neighbour (KNN)

Perhaps the easiest and most straightforward algorithm for classification is that of KNN. The algorithm involves determining the class of a sample by identifying its closest neighbours and utilizing their characteristics. The term "k-Nearest Neighbour" is used because, in many cases, it is essential to consider more than just a single neighbour (k) to classify an example (Cunningham & Jane, 2021).

Identifying the nearest neighbour is accomplished through the application of "Euclidean Distance" (Akanbi, 2014). The process in the algorithm, as well as the equation employed, are outlined as follows:

$$dx_i, x_j = \sqrt{\sum_{r-1}^{n} a_r x_i - a_r x_j{}^2}$$

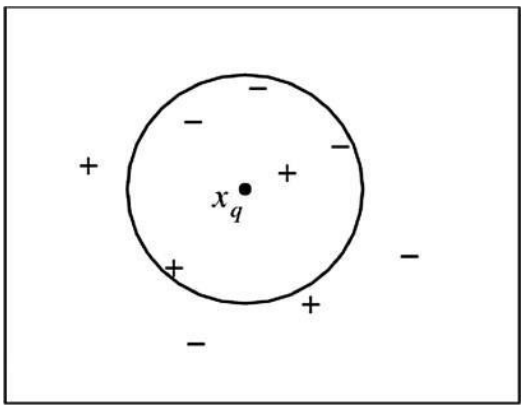


Figure 1. KNN

Source: Akanbi & Fazeldehkordi, 2014

Naïve Bayes Classifier (NBC)

NBC is a machine learning algorithm that relies on the Bayes probability theory. Through the examination of the input data corresponding to a specified set of features, the Naïve Bayes classifier can calculate class probabilities associated with a specific label (Souza et al, 2021). To classify the input data, it is necessary to calculate the probabilities associated with each existing class. The class with the highest probability is then identified as the one to which the input data belongs (Souza et al, 2021). Thus, the classification task is to locate the class with the maximum probability to occur. The Bayes theory along with the max probability equation are presented below:

$$PAB = \frac{PBAPA}{PB} \qquad a = \text{argmax}_a Pab_1 \cdots b_n$$

Source: Souza et al, 2021

Decision Trees (DT)

Decision trees serve as a versatile approach applicable to both regression and classification tasks. They stand out as one of the most employed algorithms, particularly suitable for classification tasks, offering numerous advantages over alternative classifiers. They are relatively easy to explain and interpret while requiring little effort for data preparation from the user (Benferhat & Elouedi, 2006). The structure of a typical decision tree can be observed in the figure below.

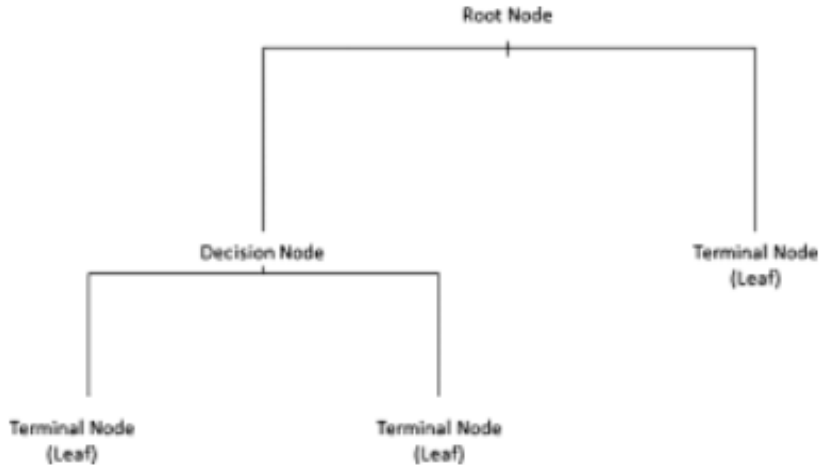


Figure 2. Decision Tree

Source: Bellini, 2019

On a tree structure the root represents the starting point, the branches represent the splits/path of the root node, while decision nodes evaluate the features to split the data and which direction to follow. Lastly, the terminal nodes represent the outcome, or while in classification task, the class label (Bellini, 2019). A useful technique that is associated with decision trees is that of pruning, which is implemented in situations that the decision tree may grow too long, so it is essential to reduce its decision nodes and avoid overfitting on the data (Bellini, 2019).

Random Forest

Frequently, it is crucial to construct more complex models by combining the predictions of several weaker models. This approach in data science is commonly referred to as the "ensemble" method. This procedure essentially combines different models that are trained to solve the same problem and get better results, making the final model more robust (Rocca, 2019).

Random Forest stands out as the most widely adopted ensemble method for classification tasks, a model that trains multiple decision trees through a process called “bootstrapping”, followed by another process known as “bagging” aggregation. Bootstrapping refers to the training of individual trees in parallel and on different subsets of the training set, utilising a diverse set of features. This process ensures that each decision tree in the random forest is unique, which consequently reduces the variance of the classifier. For the model to make the final decision, the predictions from each individual tree are aggregated leading to a more generalised method. Due to its robust nature, the random forest often surpasses other classifiers in terms of accuracy. Despite being more complex in structure compared to an individual

decision tree, the random forest is generally simpler when it comes to hyperparameter tuning (Misra & Li, 2020). A typical structure of the model can be viewed in the following image.

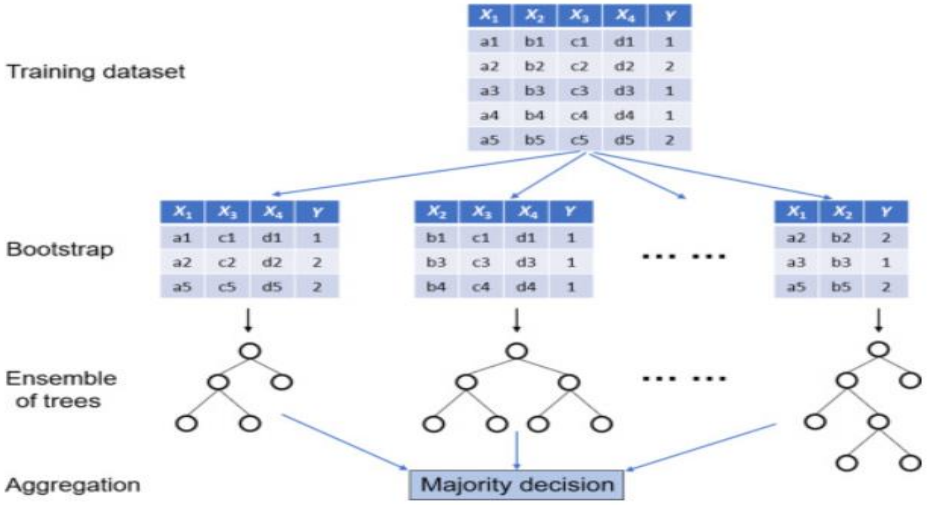


Figure 3. Random Forest

source: Misra & Li, 2020

Boosting

Another well-known ensemble method that combines numerous weak learners to create an enhanced model with improved accuracy and robustness. Comparable to the bagging technique of the random forest, boosting also involves aggregating predictions from each individual decision tree, resulting in a more generalised model. However, the primary distinction lies in the fact that, during boosting, the trees are trained sequentially, unlike the parallel training that takes place in the random forest. In this process, each model in the sequence is fitted with increased emphasis on observations in the dataset that were mishandled by the previous models. Essentially, each new model concentrates its efforts on the most challenging observations encountered so far. Consequently, by the end of the process, a robust classifier with reduced bias is obtained (Rocca, 2019). The process of boosting compared to bagging can be observed in the figure below.

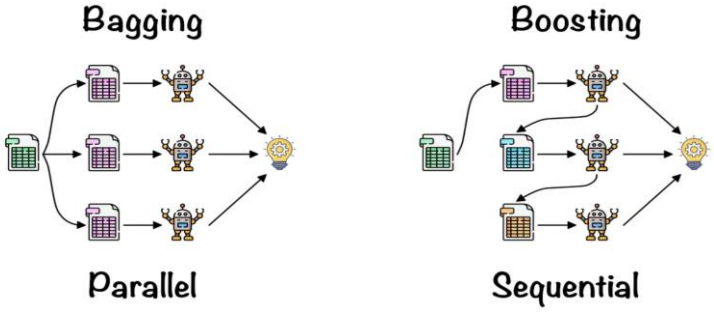


Figure 4. Boosting vs Bagging

Source: Lopez, 2021

Contrary to bagging and random forest, boosting exhibits numerous variations in the form of various algorithms. Such examples are:

- ***Adaptive Boosting***: The concept of adaptive boosting is to train multiple weak classifiers to create a robust classifier. In each iteration, the algorithm trains a weak classifier on a weighted version of the training data. The weight assigned to each training example is adjusted based on the previous weak classifiers. Adaptive boosting is implemented by the Adaboost Classifier algorithm (Polamuri, 2023).
- ***Gradient Boosting***: The process starts by training a weak learner using the initial data. It subsequently calculates the residuals, and another weak learner is fitted to the residuals. This process is repeated for a set number of iterations, where each new learner aims to minimise, the residuals left by the preceding learner. The model's prediction is the sum of all predictions from the weak learners (Polamuri, 2023).
- ***XGB (Extreme Gradient Boosting)***: An optimised version of gradient boosting, this technique has emerged as one of the most widely adopted methods for boosting procedures. This algorithm facilitates considerably faster training by enabling parallel processing and is adept at handling very large datasets with minimal requirements for data preprocessing. Like other boosting principles, XGBoost trains multiple week decision trees to create a strong model with increased accuracy and robustness (GeeksforGeeks, 2023).

Stacking

A third ensemble method is that of stacked generalization. In this approach multiple models are trained on the same dataset serving as the base estimators or “level-0 base models”. The predictions of those models are then used to train another model called as “level-1 meta model”. Hence, the output of the base models (predictions) is used as an input to the meta estimator. The main idea behind stacking generalization is to combine diverse model and combine the strengths of all those models. The meta estimator is then trained to combine those strengths for an enhanced predictive power (Brownlee, 2021). The picture below illustrates the procedure of how stacking occurs. First the training set is split into folds. K-1 folds are used to train the base models, while the validation fold is used for predictions. The predictions from all the base models are then stacked and feed the meta learner who is responsible to make the final prediction.

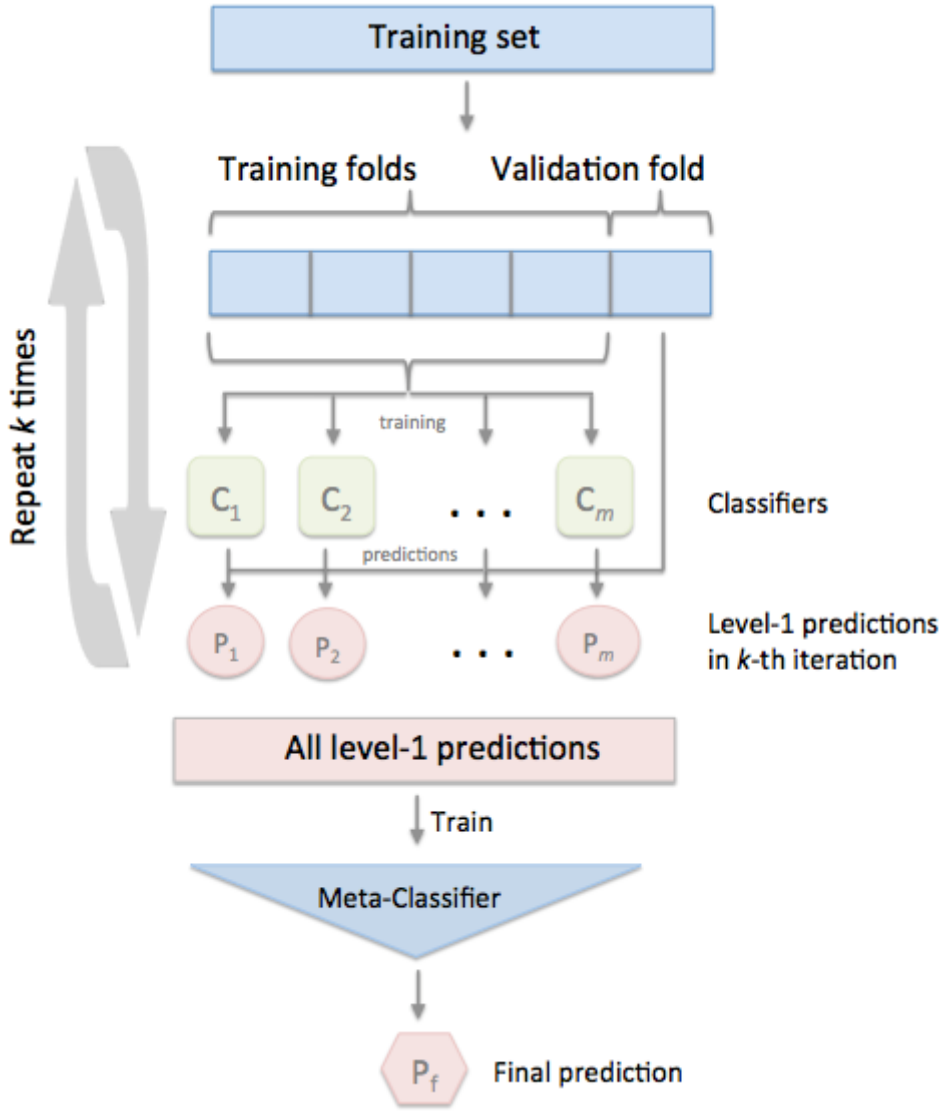


Figure 5. Stacking Method

Source: StackingCVClassifier: Stacking with cross-validation - mlxtend (rasbt.github.io)

2.4. Class imbalance and resampling strategies

On many occasions, real world data comes with an issue called class imbalance. This occurs primarily on classification tasks when the distribution of the target variable is severely skewed, meaning that a certain class is underrepresented than others. Such situations occur often in disease detection and fraud transaction datasets where there are significantly fewer observations for diseased or fraud cases. This skewness often creates challenges for the models since they become more biassed towards the majority class, resulting in poor performance for the minority classes (Yang et al, 2024). To address this situation,

resampling techniques are employed during the model training process, aiming to balance the distribution and minimise any bias in the model. The most widely used techniques are random **Undersampling** and **Oversampling**. In random undersampling, samples are deleted from the majority class until it matches the minority, though it can result in loss of information and underfitting. In contrast, with random oversampling, samples from the minority class are duplicated until it matches with the majority class. However, this considerably inflates the dataset size and increases the risk of overfitting in the training data. To mitigate such scenarios, methods like cross-validation, hyperparameter tuning, and regularisation are employed to minimise issues related to underfitting and overfitting. It is crucial though, that both of those techniques are only applied on the training set, since the test set serves as a genuine representation of the real-world problem and should always be left unaltered (Brownlee, 2021).

2.5. Evaluation metrics

The ultimate objective of any classification model is to generate predictions for new and unseen data. Hence, to assess whether a built model is effective, various evaluation methods are necessary to make a concrete decision. Those metrics are presented as followed:

Accuracy

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP = True positive; FP = False positive; TN = True negative; FN = False negative

Typically, accuracy is defined as the ratio of correct predictions to the total number of instances. However, relying solely on accuracy as a metric can be overly optimistic, especially in situations with significant class imbalance. For instance, in a scenario with 1000 transactions, of which only 100 are fraud, a classifier could predict all cases as non-fraudulent, resulting in a 90% accuracy without detecting any of the fraud cases. To address this, additional metrics are considered that can emphasising on the evaluation of minority classes.

Recall

Also referred to as sensitivity, recall is a measure of how many truly relevant results are returned. It is defined as the number of true positives over the number of true positives plus the number of false negatives.

$$\text{Recall} = \frac{tp}{tp + fn}$$

Recall is very used when you must correctly classify some event that has already occurred. It is frequently applied in scenarios such as fraud detection models or disease detection in patients. For instance, in the context of illness detection, it is crucial to identify individuals who are ill to prevent false negatives.

Precision

Precision is a measure of result relevancy. It is defined as the number of true positives over the number of true positives plus the number of false positives.

$$\text{Precision} = \frac{tp}{tp + fp}$$

Precision is widely employed in marketing campaigns, particularly in the context of marketing automation. This is because a marketing automation campaign aims to initiate an activity for a user when the system predicts a successful response. In these scenarios, low precision translates to a financial loss as it entails reaching out to potential customers who are not interested in the marketing offer.

F1 score

F1 is the harmonic mean of precision and recall. It is often preferable in situations where there is class imbalance in the dataset, while both high recall and precision are desirable. The F1 score is expressed by the following equation.

$$F_1 = \frac{2T_p}{2T_p + F_p + F_n}$$

It can also be expressed in terms of precision and recall by the following expression.

$$2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Receiver Operating Characteristic (ROC) Curve

The ROC curve depicts the balance between the true positive rate (Sensitivity or Recall) and the false positive rate (FPR) across various probability thresholds. AUC (Area Under the Curve) is a metric assessing the classifier's ability to discriminate between the two. Its value ranges from 0 to 1, with values nearing 1 signifying optimal classifier performance. While commonly employed for binary classification tasks, it is possible to generate the curve for multiclass problems as well. In this context, each class has its individual curve, while Macro-AUC represents the average of the class-specific AUC values. An example of suck curve is depicted below.

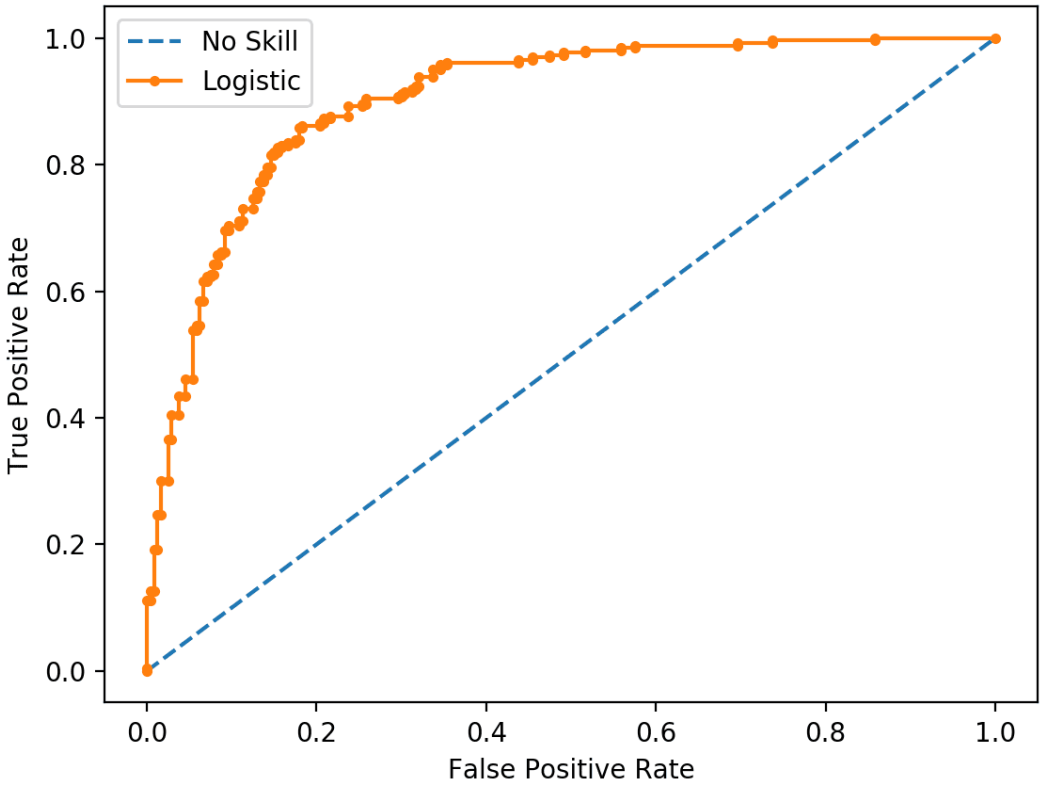


Figure 6. ROC Curve

Source: Brownlee, 2020

Average Precision (AP)

The relationship of recall and precision is depicted by the precision recall curve, which illustrates the trade-off between the two metrics for different probability thresholds. The area under the curve is known as Average Precision (AP) and is given by the following expression (Kashifi et al, 2022).

$$\mathbf{AveP} = \int_0^1 p(r)dr$$

The area under the curve, like the ROC curve, has a value ranging from 0 to 1, with higher values indicating superior classifier performance. Precision-recall curves are preferred in scenarios with severe class imbalance since AUC in those occasions can be overly optimistic. Although commonly employed for binary classification, it is feasible to calculate curves for each individual class in a one-vs-rest (OvR) scenario. The micro-AP is then computed which is the harmonic mean of the class-specific AP values. The pr curves are preferred in situations where there is a class imbalanced as they are more indicating for the performance of the minority classes. A typical pr curve is illustrated below.

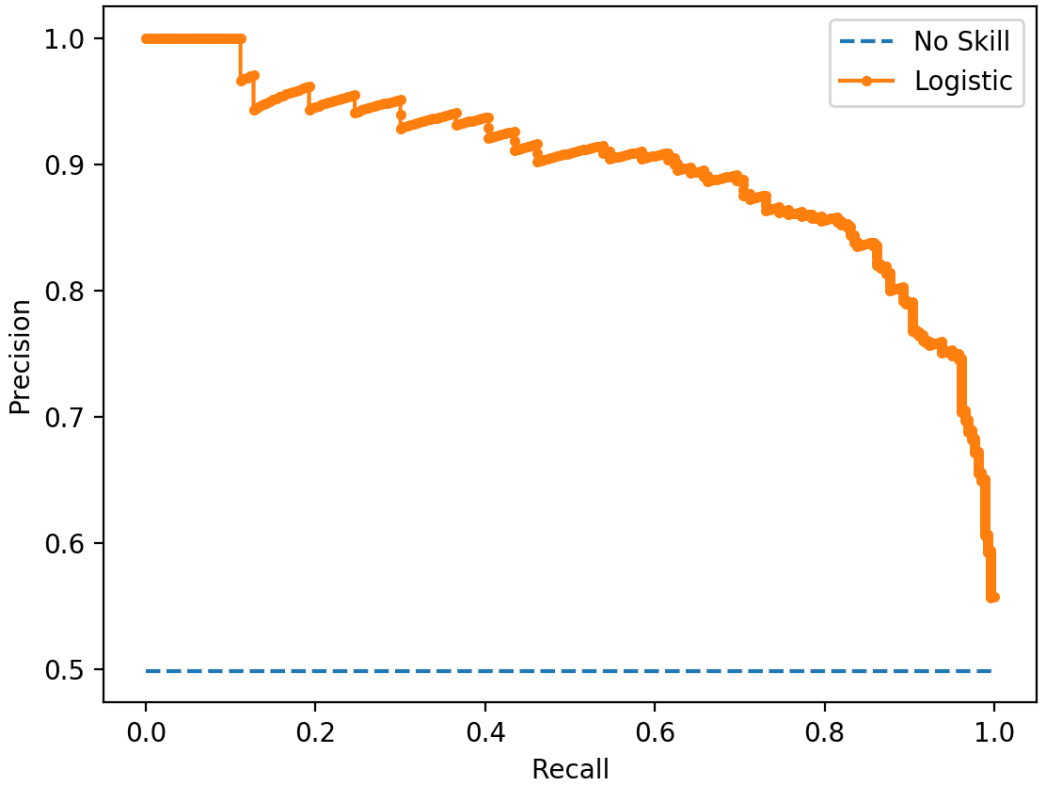


Figure 7. Precision - Recall Curve

Source: Brownlee, 2020

2.6. Machine Learning applications in transport mode choice

The adoption of machine learning in the modelling and prediction of transport mode choices has experienced a significant surge in recent times. In contrast to the prior logit models, the results indicate a notably improved ability to accurately predict transportation mode. This application is geared towards improving our understanding and predictive capabilities concerning individuals' decisions regarding the modes of transportation they choose. Numerous recent studies have sought to employ a range of machine learning models to predict transit behaviour. In 2015, Omrani conducted a study with the objective of forecasting the travel mode choices of individuals by applying machine learning techniques to national data from Luxembourg. His research outcomes revealed that artificial neural networks outperformed other alternative models in terms of predictive accuracy (Omrani, 2015). In their research conducted between 2010 and 2012 using data from the National Travel Survey in the Netherlands, Hagenauer and Helbich (2017) applied various machine learning classification models to predict travel mode choices effectively. Their results indicated that the Random Forest model outperformed the

others. Nonetheless, they noted that while trip distance emerged as the most critical predictor, the importance of variables varied among different classifiers and class labels (Hagenauer & Helbich, 2017). In a parallel study using the same data, Kashifi et al (2022) set out to also forecast transportation mode choices. They extended the previous research by incorporating additional machine learning techniques into their analysis. Their results revealed that boosting and LightGBDT exhibited superior predictive accuracy for the different classes, especially when utilizing both under and oversampling methods to address class imbalance. Furthermore, their analysis underscored that age, income, and distance were the most influential predictors in the context of transport mode prediction (Kashifi et al., 2022).

Other applications of machine learning in the prediction of transportation modes can be characterised as more tailored to specific scenarios, as they are designed for modelling particular situations. In a recent study, Bhuiya et al (2022) focused on modelling transport mode choices for individuals with limited mobility in Dhaka. Their research findings indicated that multinomial logistic regression and linear discriminant analysis models exhibited superior predictive accuracy, particularly considering a smaller dataset (Bhuiya et al., 2022). Additionally, Zhao et al (2020) investigate differences between machine learning and logit models based on trip diary recordings. Their study findings from staff and students within the University of Michigan suggest that when deciding between machine learning and logit models for transport modelling, it seems there is a trade-off between predictive accuracy and the alignment with behavioural principles (Zhao et al, 2020).

Recent research developments in this field have introduced more sophisticated approaches, such as the adoption of artificial neural networks and deep learning methods. For instance, Zhang et al. (2020) introduced a deep neural network model for classification using data from Beijing, and their results demonstrate that this network model outperforms the random forest model in predicting transportation modes (Zhang et al., 2020). Furthermore, in a study based on national travel data from the UK, Bei et al. (2023) introduce a deep neural network model that goes beyond mere travel mode prediction, also addressing the purpose of the trip. Their research indicates that the model they proposed surpasses the performance of basic multinomial logit models and single-task neural networks (Bei et al., 2023). Additionally, Wang, Mo, and Zhao (2021) present a "theory-based residual neural network" model that integrates discrete choice models with basic neural networks, using three separate survey datasets. Their results indicate that the model they propose not only achieves superior predictive accuracy but also exhibits greater resilience compared to straightforward neural networks or discrete models (Wang, Mo & Zhao, 2021).

3. METHODOLOGY

3.1. Python Libraries

Below, the Python libraries utilised for the analysis are displayed:

a) **Numpy**: It is one of the most implemented libraries, particularly in machine learning. This library is known for its support of matrices and multi-dimensional data. Numpy incorporates mathematical functions, among which the Array Interface that stands out as one of its most valuable features.

b) **Pandas**: It stands as a crucial library for data scientists, serving as an open-source machine learning library offering top notch analytical tools. It includes features such as sorting, visualisations, conversions and more.

c) **Matplotlib**: It is employed for visualising numerical data, making it a valuable tool in data analysis. It includes many unique charts such as pie charts, histograms, scatter plots and more.

d) **Scikit-learn**: A machine learning library that supports most supervised and unsupervised algorithms. The library works for many separate problems including regression, classification, clustering and more.

e) **Geopy**: A python library that enables users' identification of coordinates for addresses, cities, countries internationally. It is also used for calculating distances between different coordinates (pypi.org, 2024)

f) **Seaborn**: It is used for data visualisation, offering a high-level interface to create visually appealing and informative statistical graphics. The library is built on top of the library of Matplotlib also functioning effectively with Pandas.

3.2. Data collection

Data for Thessaloniki were collected through an online survey using the Google Docs platform. While the survey was initially created in Greek, the analysis was conducted in English. The survey was distributed both online through social media and in paper form to residents in Thessaloniki. Since the aim was to predict mode choice for commuting to work, the condition to answer the survey was for the respondent to be an active worker. The following comprises the complete list of questions posed both in Greek and English

1) Διεύθυνση Κατοικίας - Home Address
2) Διεύθυνση Εργασίας - Work Address
3) Φύλλο - Gender

4) Ηλικία - Age
5) Έχετε δίπλωμα για οποιαδήποτε από τα παρακάτω: Αυτοκινητο, Μηχανη, Τίποτα - Do you have a license for any of the following: Car, Motorcycle, Nothing
6) Διαθέτετε κάποιο από τα παρακάτω: Ποδήλατο, πατίνι, τίποτα - Do you have any of the following: Bicycle, skates, nothing
7) Από πόσα άτομα αποτελείται η οικογένειά σας - How many people does your family consist of ?
8) Πόσα ιδιωτικά οχήματα διαθέτετε στην οικογένειά σας - How many private vehicles do you have in your family?
9) Μηνιαίο εισόδημα - Monthly Income
10) **Με ποιον τρόπο πηγαίνετε συνήθως στην εργασία σας - Mode for commuting to work (target variable)**
11) Αναφέρετε πόσα λεπτά στο περίπου χρειάζεται για να μεταβείτε από το σπίτι σας προς τον χώρο εργασίας σας - Indicate how many minutes approximately it takes you to travel from your home to your workplace.
12) Ποιές ώρες πηγαίνετε συνήθως στην εργασία σας - What time do you usually go to work?

Likert Based questions

Do you think the following factors affect your commute?

Use scale where:

1-Strongly disagree , 2-Disagree , 3-Neutral , 4-Agree , 5-Strongly agree

13) Άνεση Μετακίνησης - Convenience
14) Κόστος Μετακίνησης - Cost
15) Ασφάλεια μετακίνησης - Safety
16) Προστασία του περιβάλλοντος - Environmental concerns
17) Σωματική άσκηση και υγεία - Physical exercise and health
18) Καιρικές συνθήκες - Weather conditions
19) Διαθεσιμότητα χώρου στάθμευσης - Parking availability

The survey remained accessible from November 15 to January 15. A total of 409 samples were gathered, and after eliminating blank or erroneous responses as well as samples with unclear home or work addresses, 383 were deemed valid for subsequent analysis. The sample was visualized in a Map of Thessaloniki using the home addresses and the use of geopy library in python. The library calculates the coordinates of the home address and creates a map with markers based on those coords. The code used along with the sample visualization are illustrated below.

```
addresses = df['Home_address'].tolist()

# Geocode addresses to obtain latitude and longitude
geolocator = Nominatim(user_agent="address_visualization")
locations = []

for address in addresses:
    location = geolocator.geocode(address)
    if location is not None:
        locations.append(location)

# Create a DataFrame with latitude and longitude
coordinates = [(location.latitude, location.longitude) for location in locations]
df_coordinates = pd.DataFrame(coordinates, columns=['Latitude', 'Longitude'])

# Create a Folium map centered at the mean coordinates
map_center = [df_coordinates['Latitude'].mean(), df_coordinates['Longitude'].mean()]
map_object = folium.Map(location=map_center, zoom_start=12)

# Add markers for each address
for index, row in df_coordinates.iterrows():
    folium.Marker(location=(row['Latitude'], row['Longitude'])).add_to(map_object)
```

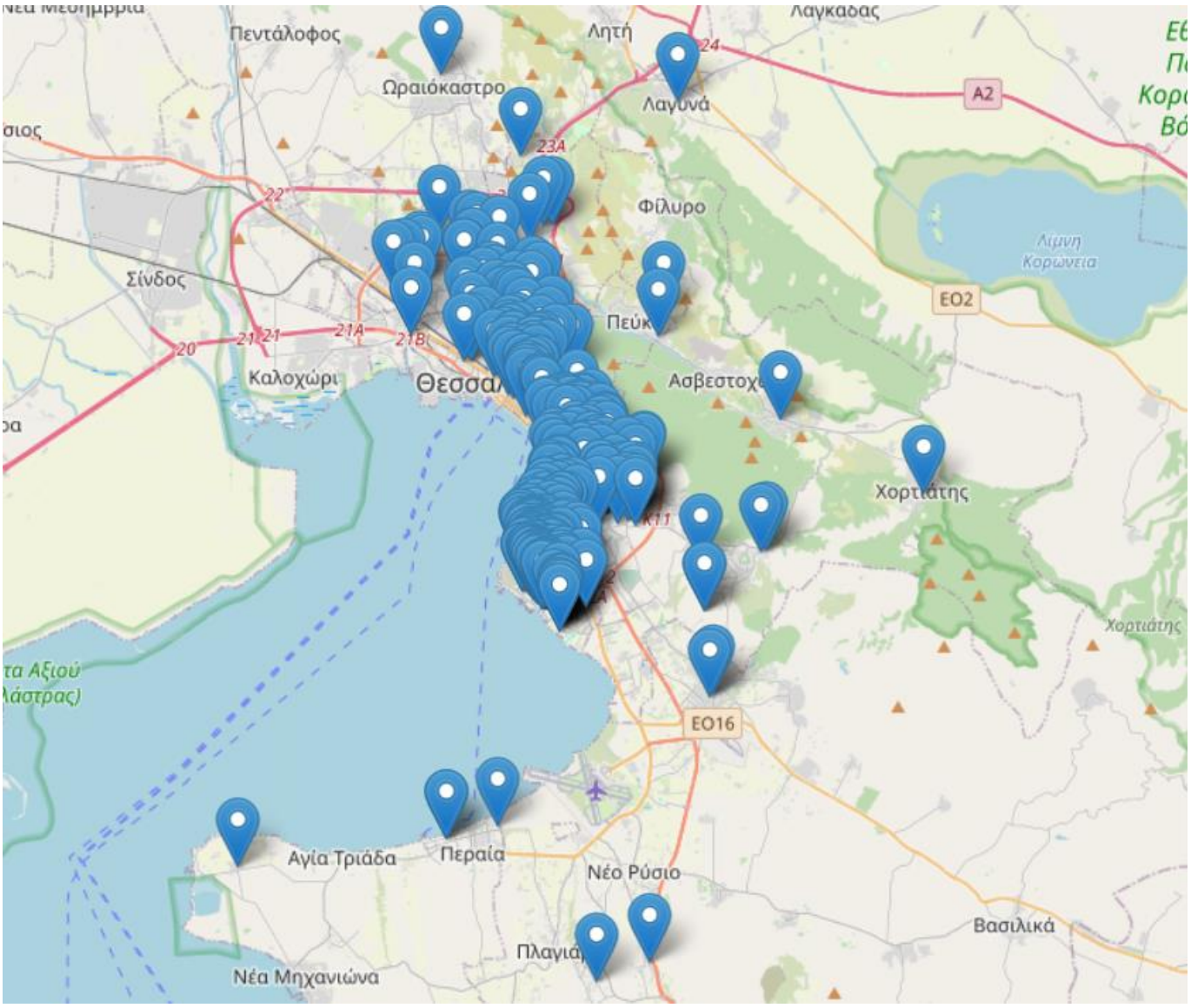


Figure 8. Thessaloniki sample visualization

As identified in the picture above, the sample was gathered across various locations in the city of Thessaloniki, so it could be representative of commuters in various areas.

Each question of the survey corresponds to a feature in the resulting dataset. Additional columns were generated for checkbox questions like 5, indicating binary values of either yes or no. For instance, the possible answers in question 5 are Car, Motorcycle or Nothing. Two columns were created: Car_licence and Motor_licence that store whether the respondent has a licence for those or not. The same procedure was implemented also for question 6, creating two new features: Bike_access and Skate_access. Additionally, two distance-related features were generated using the Home address and Work address. More specifically, distance in kilometres was manually computed between home and work address using Google Maps. The shortest path from home to work address was picked for every instance. An illustration of the procedure is presented in the figure below:

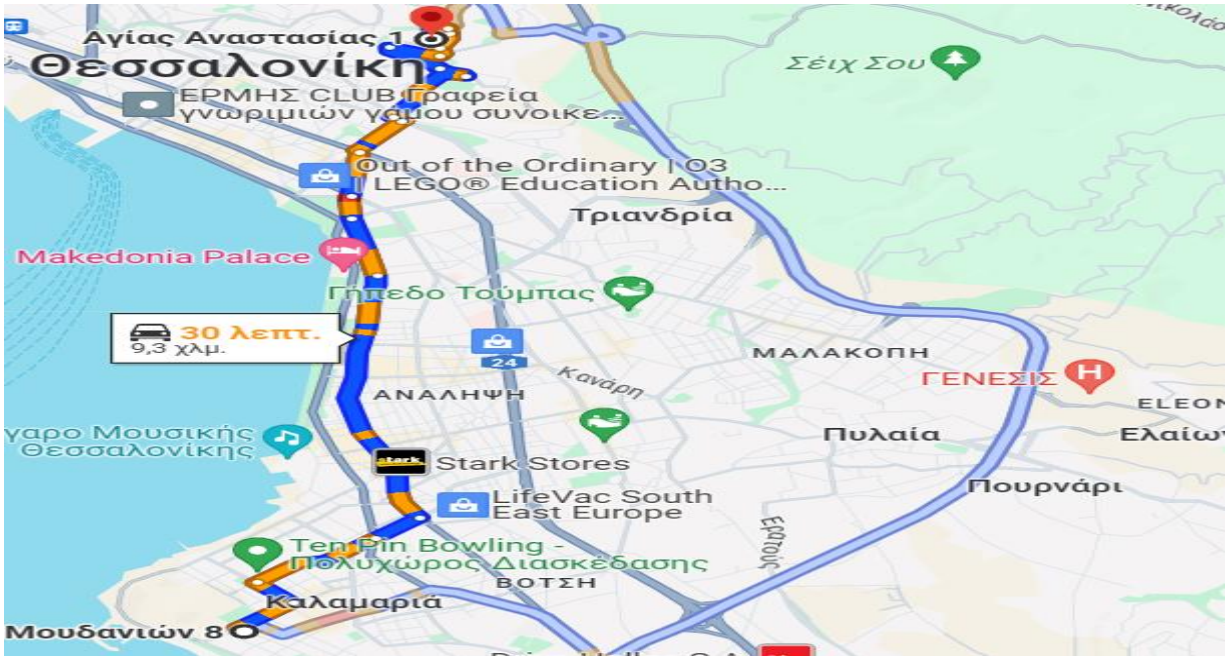


Figure 9. Google Maps Distance calculation

Also, the corresponding time for each distance was used to validate the responses from the survey in question 11 and the time (in minutes) required for commute. Specifically, respondents were requested to indicate the duration of their commute to work in minutes. In cases where there was a substantial difference between their provided answers and the calculated time in Google Maps, the corresponding time from Google Maps was adopted. For example, if a respondent stated a 40-minute commute, but Google Maps suggested lower than 20, the Maps estimate was used. In contrast, if a respondent implied a 30-minute commute and Google Maps indicated 33 or 34 minutes, the respondent's provided time was selected. This procedure was manually executed for all 383 samples.

The second distance-related feature was calculated in a more automated way using the Python library of Geopy. The library provides access in geocoding enabling users to find the geographic coordinates of a location based on the provided address. Based on those coordinates Geopy can calculate the

distance between home and work address in a completely automated way. Though, the main difference from Google Maps, is that the corresponding distance is a geodesic one. Geodesic distance is the shortest distance between two points but on a curved surface like that of the earth (Kettle, 2014). This can be better understood for distances using the airplane, where in a 2D map the shortest path between two points seems to be a straight line, but the actual shortest path is the geodesic one because of the curved surface of the earth. This example is demonstrated in the figure below with the geodesic distance represented by the blue line.

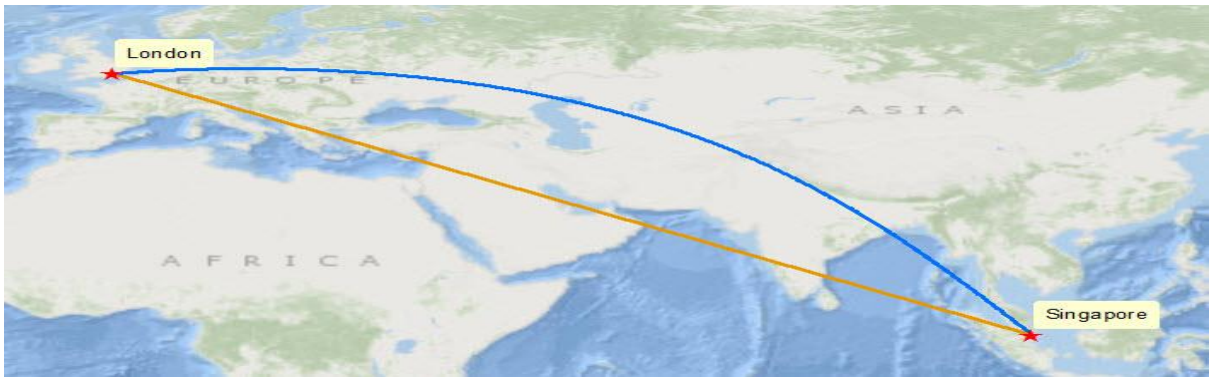


Figure 10. Geodesic Distance example

Source: Kettle, S., 2014

The purpose of calculating two types of distances is to use them individually as features in classification models and evaluate whether there are significant differences in the model outcomes depending on the use of separate distance metrics. The resulting dataset comprises 383 samples and 24 features. The analysis procedure that was followed for the dataset is illustrated in the figure below:

3.3. Analysis procedure

The figure below illustrates the analysis procedure for the dataset. Initially, the dataset was explored for erroneous answers and outliers. Afterwards, an exploratory data analysis was executed to better understand the dataset and identify patterns. The data was the preprocessed so the models could be applied. Preprocessing includes encoding into numeric form, setting up train and test sets and normalization of data. Four models were applied: Decision Tree, Random Forest, XGBoost and a Stacking model. Different iterations were used to build the models such as feature reduction, swapping the distance metrics and parameter tuning. The model predictive performance was evaluated based on overall accuracy, recall, precision and auc score. The model with the best overall performance on those metrics was selected as the most appropriate for predictions on new unseen data.

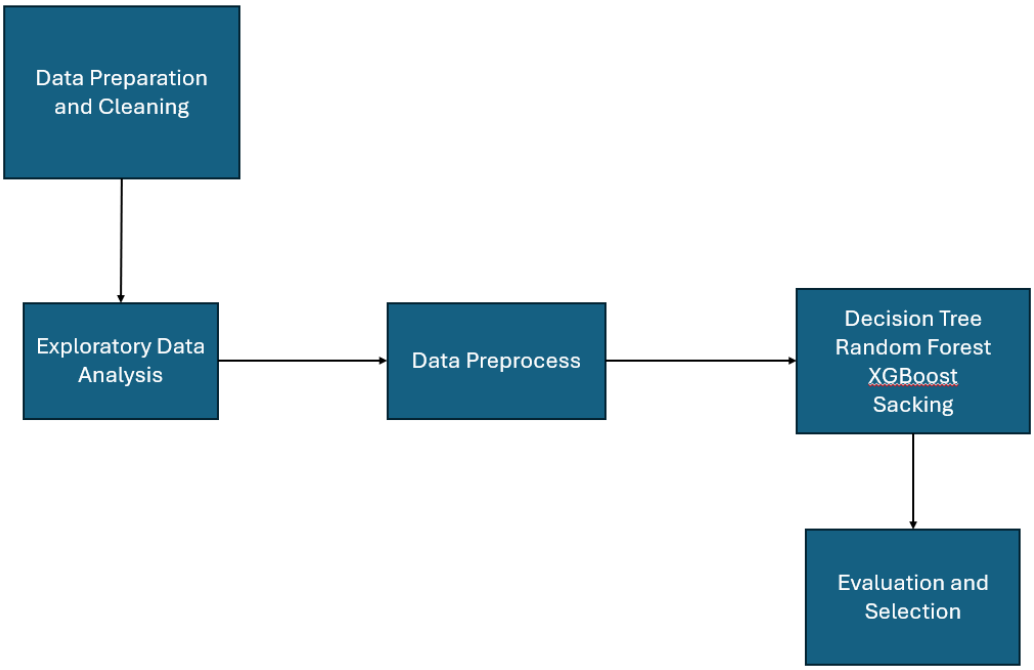


Figure 11. Analysis procedure

4. RESULTS

4.1. Results for Thessaloniki

This section of the thesis will present the findings related to Thessaloniki and commuting behaviour. The initial part covers data preparation and cleaning, followed by exploratory data analysis and data preprocessing. The final segment focuses on model application, mode classification, and the selection of the best model.

4.1.1. Data preparation and cleaning

Out of the 409 samples obtained from the online survey, blank responses were excluded initially. Subsequently, responses with errors in home or work addresses (those unidentifiable in Google Maps) were eliminated. Lastly, a few samples indicated respondents working from home or concealing either of their addresses; these samples were also excluded. Consequently, from the initial 409 samples, 383 were deemed valid. Furthermore, as detailed in the methodology section, additional features were generated for the two checkbox questions, considering only "yes" or "no" values. Following the creation of these features, the data were imported into Python for further preparation and cleaning. Below are the columns of the dataset as imported into the virtual environment:

The column names, excluding those derived from checkboxes, essentially correspond to the questions posed in the online survey. For better management during the analysis phase, it was necessary to rename the columns. The features used are illustrated below.

Table 1. List of features

Columns	Type	Description
Home_address	object	Home address of the respondent
Work_address	object	Work address of the respondent
Gender	Categorical-Binary	Gender of the respondent
Age	Categorical	Age group of the respondent
Driver_licence	Categorical - Binary	Access to a driver licence?
Motor_licence	Categorical - Binary	Access to a motor licence?
Bike	Categorical - Binary	Access to a driver bike?
Skate	Categorical - Binary	Access to a skate?
Hsize	Numeric - Discrete	Household size
Vehicles	Numeric - Discrete	Number of private vehicles in household
Income	Categorical	Income group of respondents
Mode	Categorical	Mode for commuting to work (Private vehicle, bus, walk)
Time	Numeric - Continuous	Minutes required for transit to work
Depart_time	Categorical	Time of departure for commuting to work
Convenience	Categorical - Ordinal	Degree of influence of convenience (1 to 5)
Cost	Categorical - Ordinal	Degree of influence of cost (1 to 5)
Safety	Categorical - Ordinal	Degree of influence of safety (1 to 5)
Environment	Categorical - Ordinal	Degree of influence of environmental concerns (1 to 5)
Health	Categorical - Ordinal	Degree of influence of exercise and health (1 to 5)
Weather	Categorical - Ordinal	Degree of influence of weather (1 to 5)
Parking	Categorical - Ordinal	Degree of influence of parking availability (1 to 5)
Distance	Numeric - Continuous	Travel distance (kms) between home and work address

In total, there are 24 features, comprising 4 numeric, 16 categorical, and 2 object features. The next steps involved checking for null values using the **df.isnull().any()** command and identifying duplicates with **df.duplicated().sum()**. Given that blank samples were eliminated prior to importing into Python, there were no instances of nulls or duplicates.

Furthermore, erroneous samples were also removed. There were cases where respondents indicated commuting to work by private vehicle without possessing a licence for either of those. Also, it was examined whether there are instances of respondents commuting to work with a private vehicle without having any in their household, though no such answer was found.

Afterwards, a lone sample was excluded, where the respondent suggested a trip distance exceeding 80 kilometres, constituting an outlier within the dataset. The last step was to calculate the geodesic distance using the geopy library in python. The code is illustrated below.

```
# Create a geolocator instance with a unique user agent
geolocator = Nominatim(user_agent="MyGeocodingApp_Marios_Melachroinos_v2123242526")


# Function to get coordinates for an address
def get_coordinates(address):
    location = geolocator.geocode(address)
    if location:
        return location.latitude, location.longitude
    else:
        return None

# Function to calculate distance between home and work coordinates
def calculate_distance(row):
    home_coords = get_coordinates(row['Home_address'])
    work_coords = get_coordinates(row['Work_address'])

    if home_coords and work_coords:
        return geodesic(home_coords, work_coords).kilometers
    else:
        return None

# Apply the calculate_distance function to each row in the DataFrame
df['Geodesic_distance'] = df.apply(calculate_distance, axis=1)
```

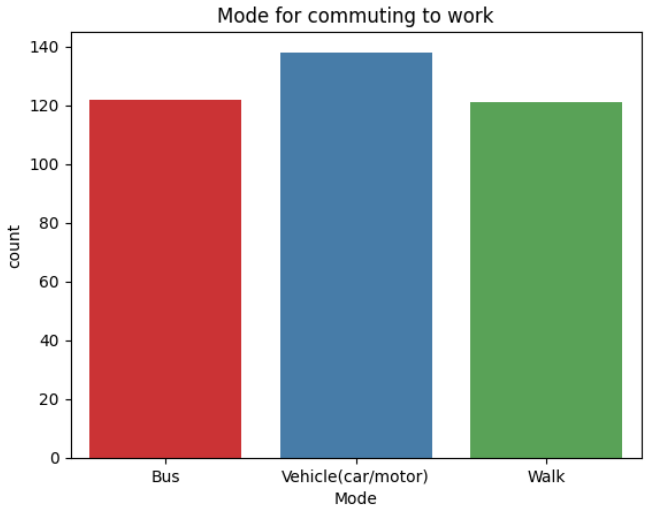
There were samples where the process could not identify distances between addresses, so NaN values returned for those cases as shown below.

Distance	Geodesic_distance
8.00	7.157182
7.60	13.814558
6.60	6.063515
0.35	NaN

Those NaN values were replaced with the corresponding Distance (that was calculated manually). For instance, the above NaN value was replaced with 0.35. The same procedure was followed for all instances with NaN values. Consequently, 381 valid options were retained to proceed with the exploratory data analysis phase. The final dataset consists of 381 samples and 21 features.

4.1.2. Exploratory Data Analysis

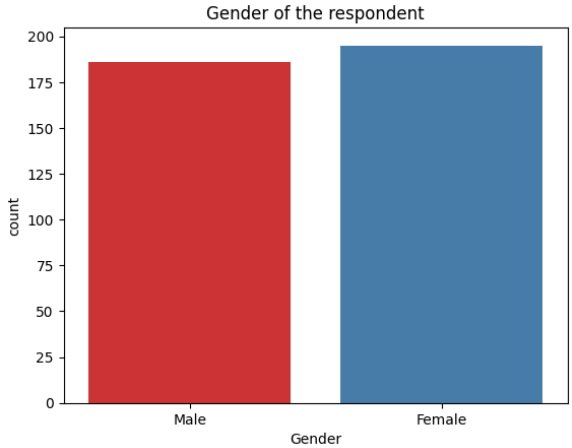
In this section of the Thesis the descriptive statistics along with visualisation for each feature will be presented to better understand the dataset. The first feature is about the target variable and the commuting mode to work for the respondents. It is evident that there is a balance among private vehicle, walk and bus users with over 100 observations for each. Note that private vehicle includes both commuters via car and motorcycle. Displayed below is the count plot of the feature, accompanied by the corresponding percentages for each mode.



Mode	Percentage
Bus	32.02%
Walk	31.75%
Vehicle (car/motor)	36.22%

Figure 12 Transport Mode

The following feature refers to the gender of the respondents. It is apparent that there is a relatively even distribution among the responses. To be more specific, there are slightly over 175 observations for both men and women, with women having a slim numerical edge (51.18%). The grouped plot also illustrates that a higher proportion of women use the bus compared to men. In contrast, men tend to use private vehicles slightly more than women.



Gender	Percentage
Women	51.18 %
Men	48.82 %

Figure 13. Gender

The subsequent set of features pertains to whether the respondent possesses driving licences. The options for car and motor licences only allow for "yes" or "no" responses. The plots make it apparent that most respondents possess a car licence compared to motor licence users.

	Driver licence (car)	Motor licence
Yes	**68,24 %**	**84.25 %**
No	**31.75 %**	**15.74 %**

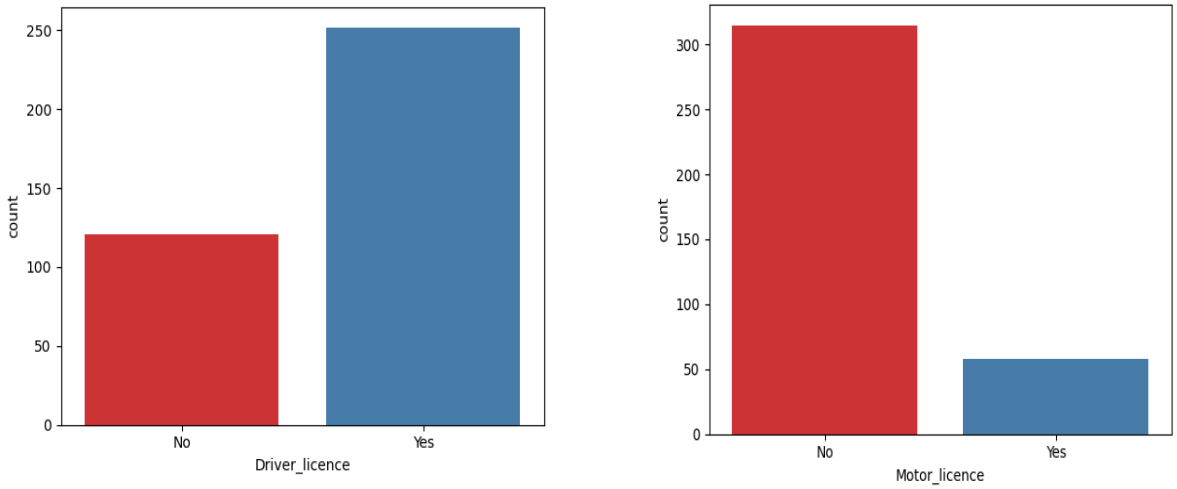


Figure 14. Driver and Motor licence

The following group of features pertains to whether the respondent has access to bike, or skate. Like the preceding set, these features only permit "yes" or "no" as possible answers. In both situation most of the respondents imply not possessing any of those.

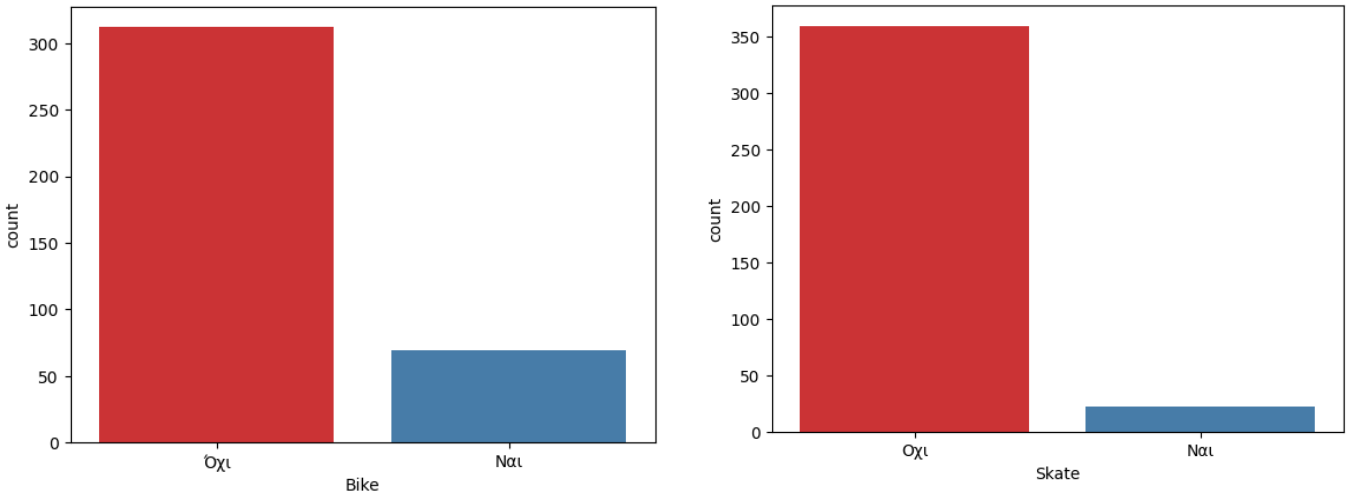
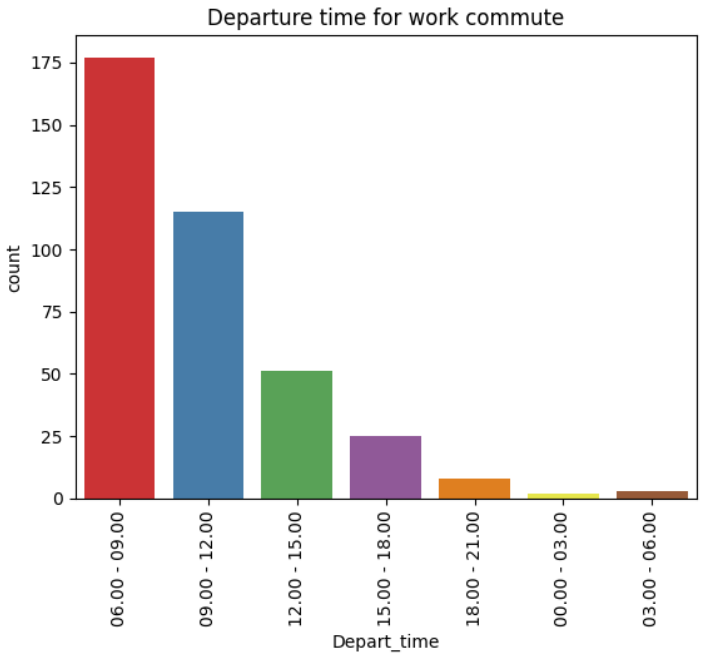


Figure 15. Skate and Bike access

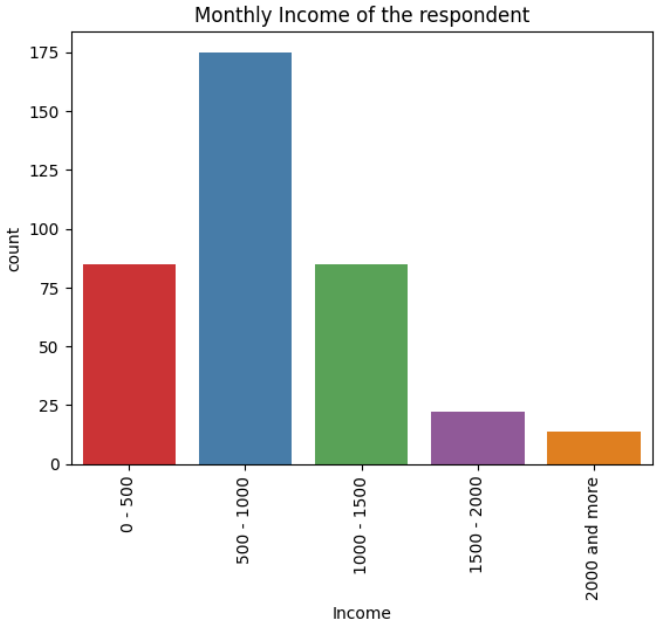
The next feature concerns the departure time of the respondent for their work commute. As depicted in the plot below, it is apparent that most respondents commute to work during the morning peak hours between 06:00 and 12:00. There are also a few instances where respondents commute during the evening hours between 12:00 and 18:00, while the minority proportion commutes to work at off hours during 18:00 - 21.00 and 00.00 - 06.00 past midnight. There are no instances where respondents depart between 21.00 and 00.00.



Depart time	Percentage
06.00-09.00	46.45 %
09.00-12.00	30.18 %
12.00-15.00	13.39 %
15.00-18.00	6.56 %
18.00-21.00	2.09 %
00.00-03.00	0.78 %
03.00-06.00	0.52 %

Figure 16. Departure time

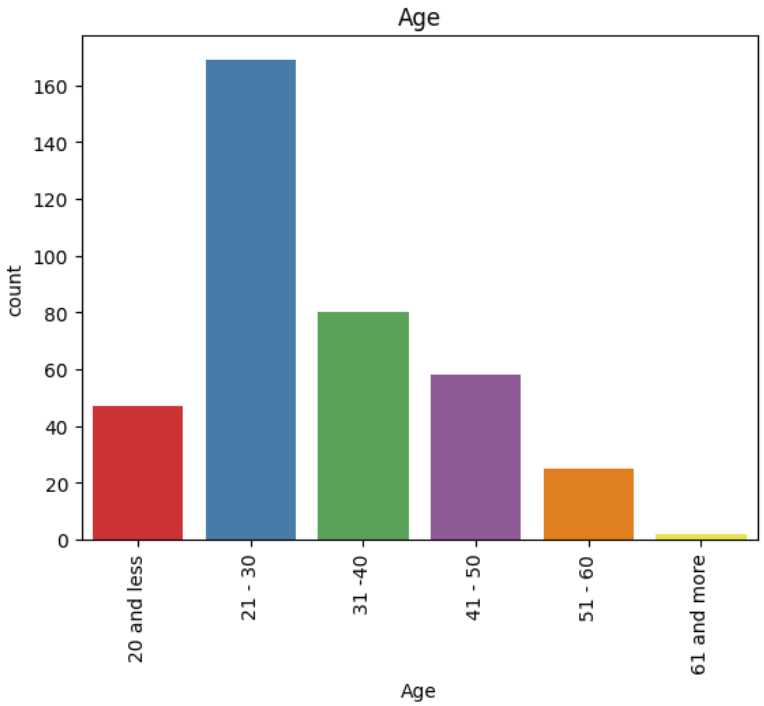
The following question refers to the monthly income of the respondent. As depicted in the graph, most respondents have a monthly income between 500 and 1000£, followed by those with incomes between 0 and 500£. Additionally, there is a substantial proportion of respondents with an income exceeding the 1000£ mark, with a few observations even surpassing 2000£.



Income	Percentage
0 - 500	22.30 %
500 - 1000	46.93 %
1000 - 1500	22.30%
1500 - 2000	5.74 %
2000 and more	3.67 %

Figure 17. Income

The next feature pertains to the age group of the respondent. The vast majority fall within the 21 to 30 years old category, followed by those aged between 31 and 40. For the remaining age groups, the observations are relatively more balanced. The sole exception is the age group of 61 and above, which has fewer than 5 observations.



Age	Percentage
Below 20	12.33 %
21 - 30	44.35%
31 - 40	20.99 %
41 - 50	15.23 %
51 - 60	6.56 %
More than 61	0.52%

Figure 18. Age

The following feature is about the household size of the respondent. The overwhelming majority indicates residing in households comprising between 2 and 5 members. In addition to this, a minority of respondents suggests living in larger families with 6 or more members, while a small proportion mention being part of a single-member family.

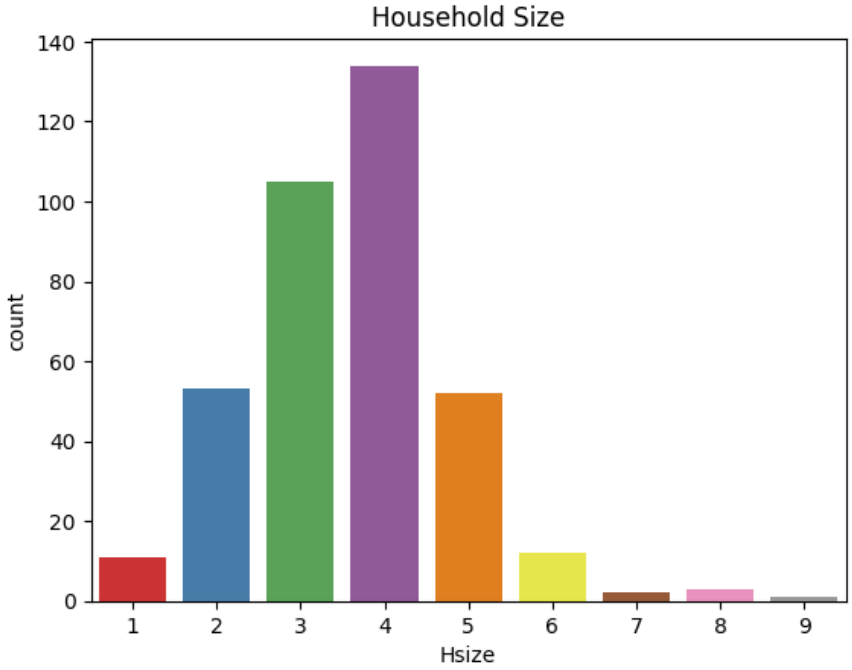
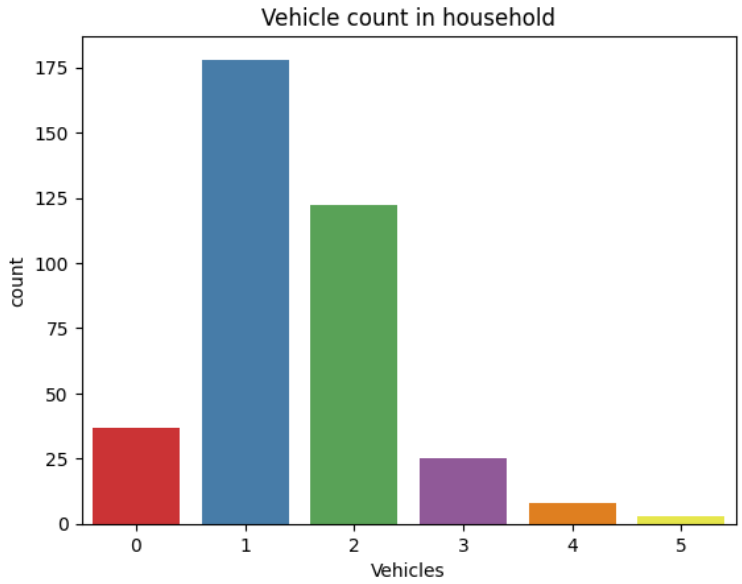


Figure 19. Household Size

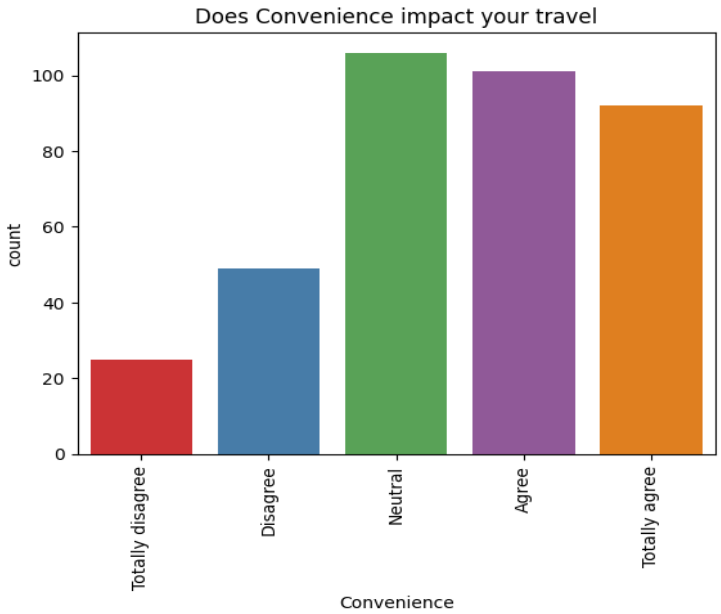
The subsequent feature concerns the number of vehicles in the households of the respondents. Most respondents suggest living in a household with 1 or 2 private vehicles. Additionally, a minority of respondents mentions residing in a house with no vehicles at all, while a few observations suggest households with 4 or 5 vehicles.



Number of vehicles	Percentage
0	9.92 %
1	47.72 %
2	32.71 %
3	6.70 %
4	2.14 %
5	0.80 %

Figure 20. Number of vehicles

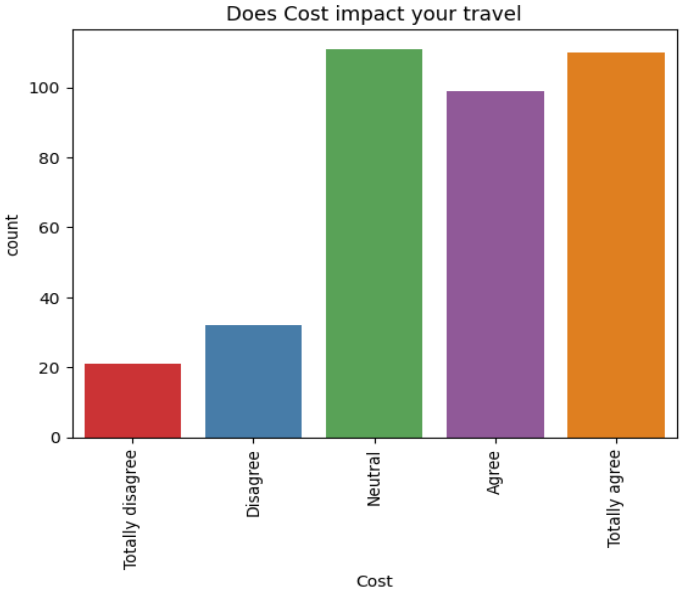
The next series of questions addresses the impact of specific factors on the commuting behaviour of the respondent. Starting with convenience as an influencing factor, a significant number of respondents expressed either "agree" or "totally agree" that the convenience of transportation affects their commuting behaviour to work. Additionally, there is a substantial proportion of respondents indicating a "neutral" response to the impact of convenience. Fewer observations are categorised as "disagree" or "totally disagree."



Impact of Convenience	Percentage
Totally Disagree	6.56 %
Disagree	12.86 %
Neutral	28.34 %
Agree	27.29 %
Totally Agree	24.93 %

Figure 21. Convenience

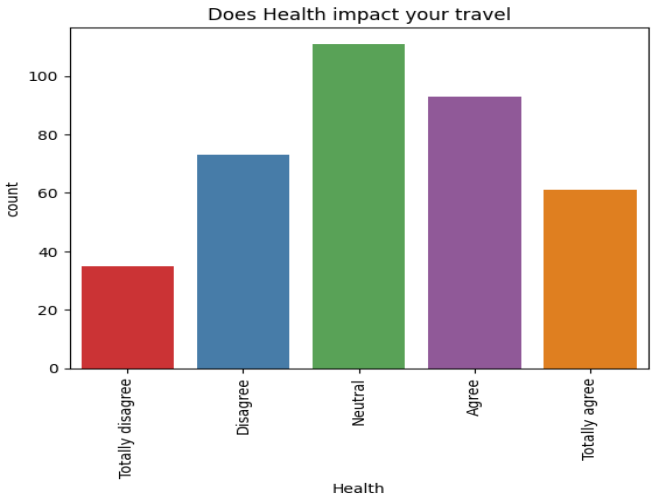
The next feature refers to the impact of transportation cost in commuting behaviour. Most of the respondents also expressed “Agree” and “Totally Agree” that cost is a significant factor influencing their transits to work. Additionally, a significant portion expressed a "neutral" position regarding this aspect. In contrast, only a small number of responses suggested that cost does not impact their commuting behaviour, as reflected in the "Disagree" and "Totally Disagree" categories.



Impact of Cost	Percentage
Totally Disagree	5.51 %
Disagree	8.39 %
Neutral	29.92 %
Agree	25.98 %
Totally Agree	30.18 %

Figure 22. Cost

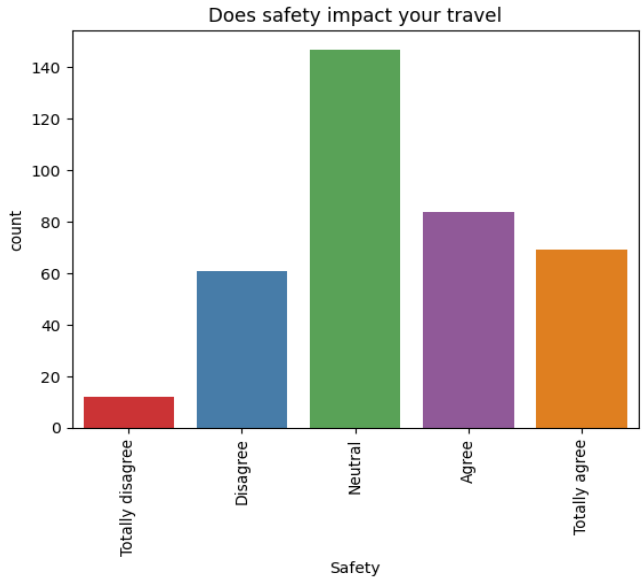
The following factor is about physical exercise and health. Specifically, it refers to how physical exercise such as walking and consequently health concerns in general affect the commuter’s transit to work. The results are quite balanced with respondents implying both negative and positive categories. Also, a significant number of respondents also indicated a “neutral” position regarding health impact.



Impact of Health	Percentage
Totally Disagree	9.44 %
Disagree	19.42 %
Neutral	29.92 %
Agree	24.93 %
Totally Agree	16.27 %

Figure 23. Physical activity and health

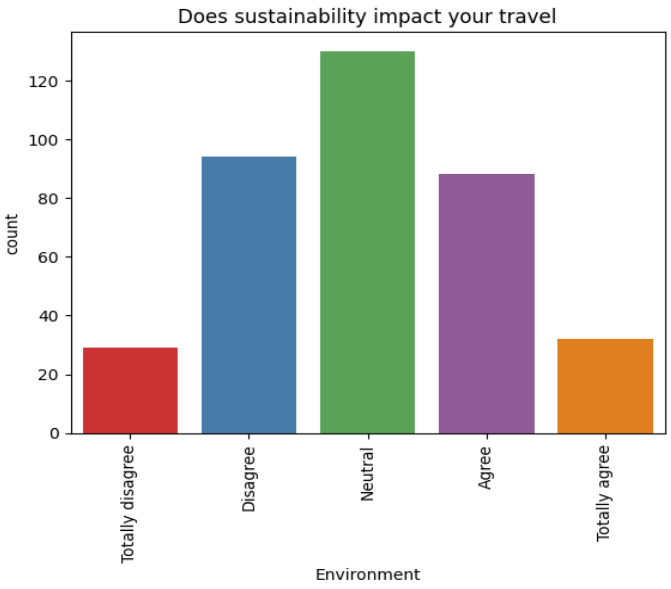
The next feature pertains to the factor of transportation safety. Once again, most respondents express a neutral stance regarding the impact of safety on their commuting behaviour to work. Additionally, a significant proportion reacts positively to the importance of this factor. A smaller number of respondents indicate that safety does not influence their transit behaviour.



Impact of Safety	Percentage
Totally Disagree	3.41 %
Disagree	16.01 %
Neutral	39.10 %
Agree	22.57 %
Totally Agree	18,89 %

Figure 24. Safety

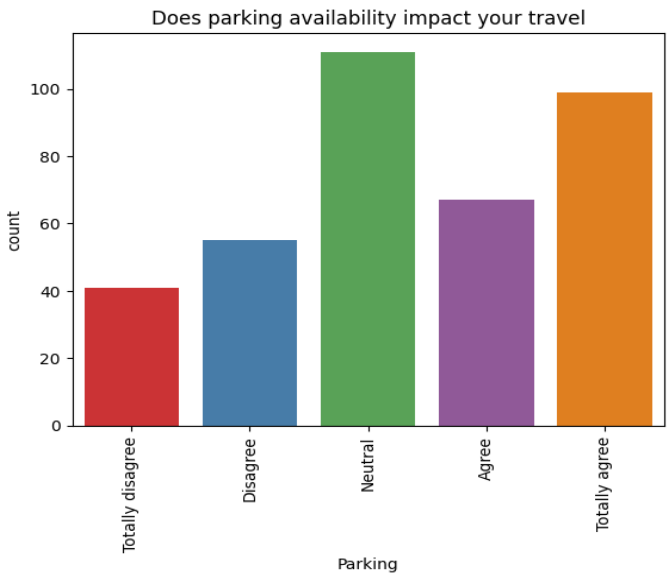
The subsequent feature focuses on the influence of environmental concerns. In a manner akin to health and physical exercise, respondents exhibit a balanced stance, with most observations equally distributed across positive and negative categories. Once more, many responses align with a neutral stance for this specific factor.



Impact of Environmental concerns	Percentage
Totally Disagree	7.87 %
Disagree	24.93%
Neutral	34.64 %
Agree	23.88 %
Totally Agree	8.66 %

Figure 25. Environmental concerns

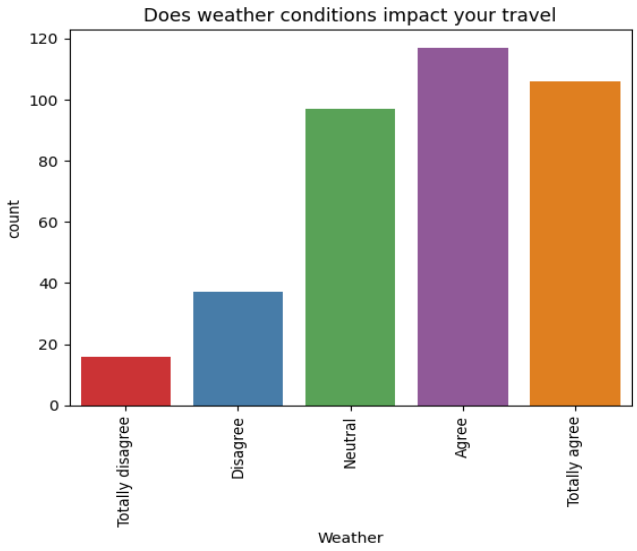
Transitioning to the next factor, it concerns the impact of parking availability on commuting behaviour to work. In this context, numerous observations fall into the positive categories, particularly the "totally agree" option, with the "neutral" category being the second most common. The negative categories encompass fewer observations regarding the impact of this factor.



Impact of parking availability	Percentage
Totally Disagree	11.02 %
Disagree	14.69 %
Neutral	29.39 %
Agree	18.37 %
Totally Agree	26.50 %

Figure 26. Parking availability

The final factor concerns the impact of weather conditions on commuting to work. In this aspect, there is a notable contrast between positive and negative responses, with most respondents falling into the "agree" and "totally agree" categories. A substantial proportion also indicated a neutral stance on the factor, while both the "disagree" and "totally disagree" categories represent vast minorities.



Impact of weather	Percentage
Totally Disagree	4.29 %
Disagree	14.74 %
Neutral	29.76 %
Agree	17.96 %
Totally Agree	26.54 %

Figure 27. Weather impact

The subsequent set of features pertains to the distance metrics calculated between the home and work addresses. The "Distance" feature signifies the optimal travel distance manually obtained from Google Maps, whereas the "Geodesic Distance" represents the distance automatically calculated by the Geopy library in Python. Both distances are expressed in kilometres. The histograms and boxplots for both metrics, along with their descriptive statistics, are depicted below.

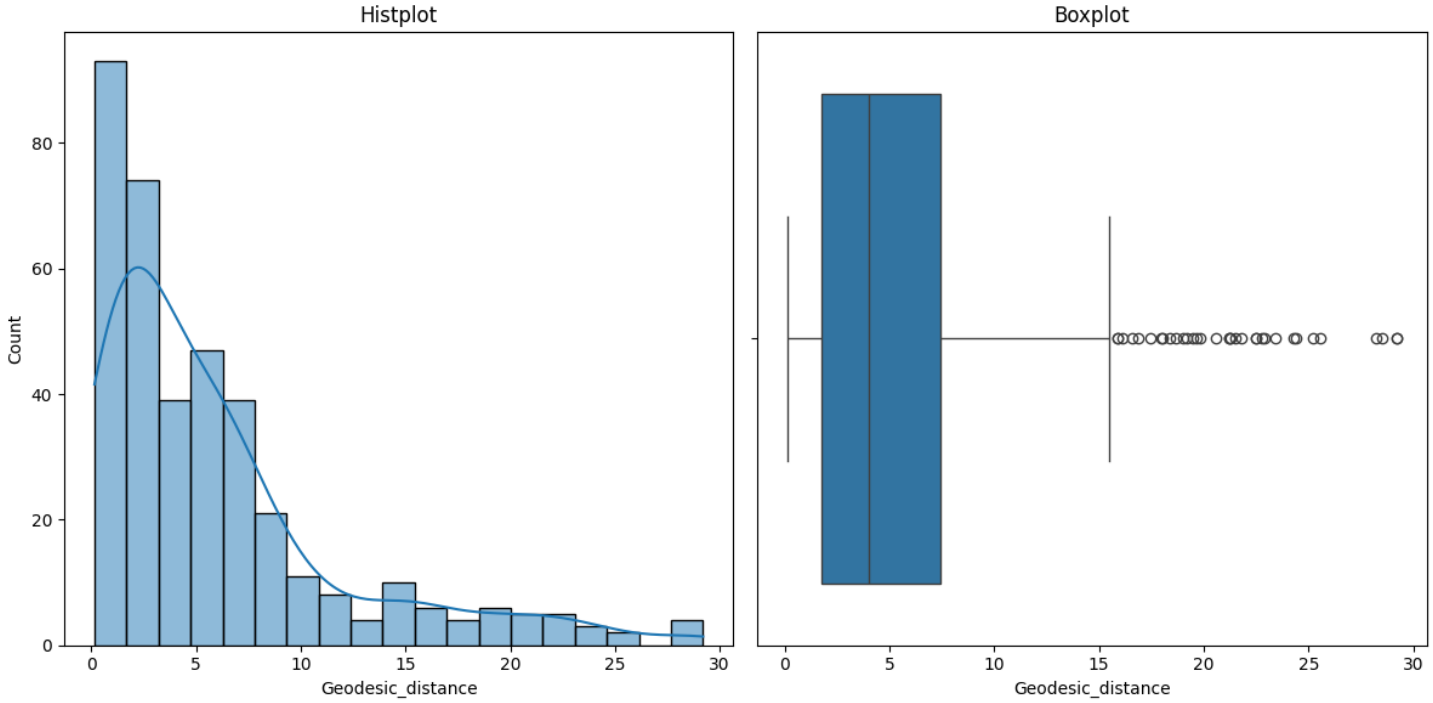


Figure 28. Geodesic Distance plots

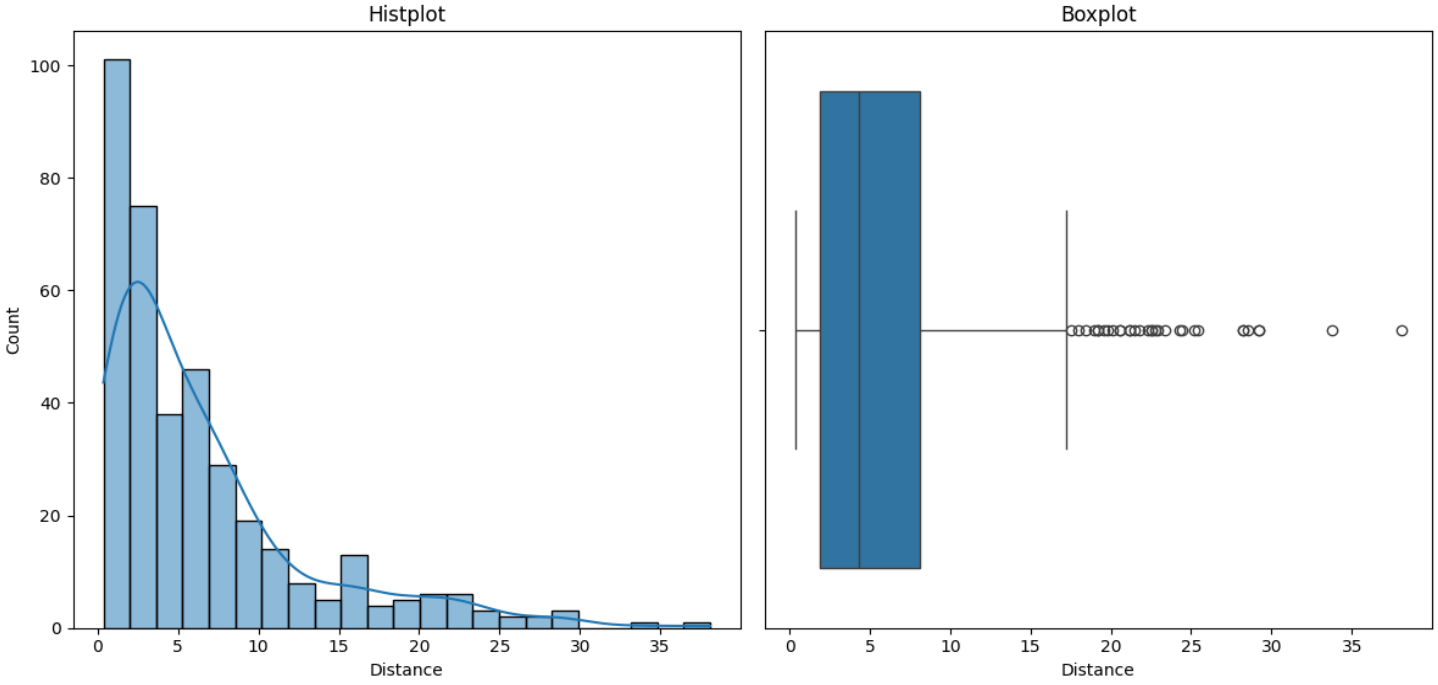


Figure 29. Distance plots

Table 2. Distance and Geodesic Distance descriptives

	Distance	Geodesic Distance
Mean	**6.48**	**5.97**
std	**6.51**	**6.01**
min	**0.35**	**0.16**
25 %	**1.9**	**1.73**
50 %	**4.3**	**4.04**
75 %	**8.1**	**7.41**
max	**38.1**	**29.2**

The graphs reveal that most observations are concentrated in the initial classes, ranging from 0.1 to 8 kilometres, regardless of the distance metric calculated. Furthermore, both distributions exhibit positive skewness while all the descriptive statistics for geodesic distance appear slightly lower when compared to optimal travel distance. The graphs below show the density of each mode for both distance metrics. Density represents the concentration of data points for a continuous variable in a normalized version over count plots. As illustrated on the plots below the concentration of data for walk is mostly within 0 to 5 kilometres while for bus is within 0 and 10 kilometres. For private vehicle the distribution is more spread across 0 and 35 kilometres.

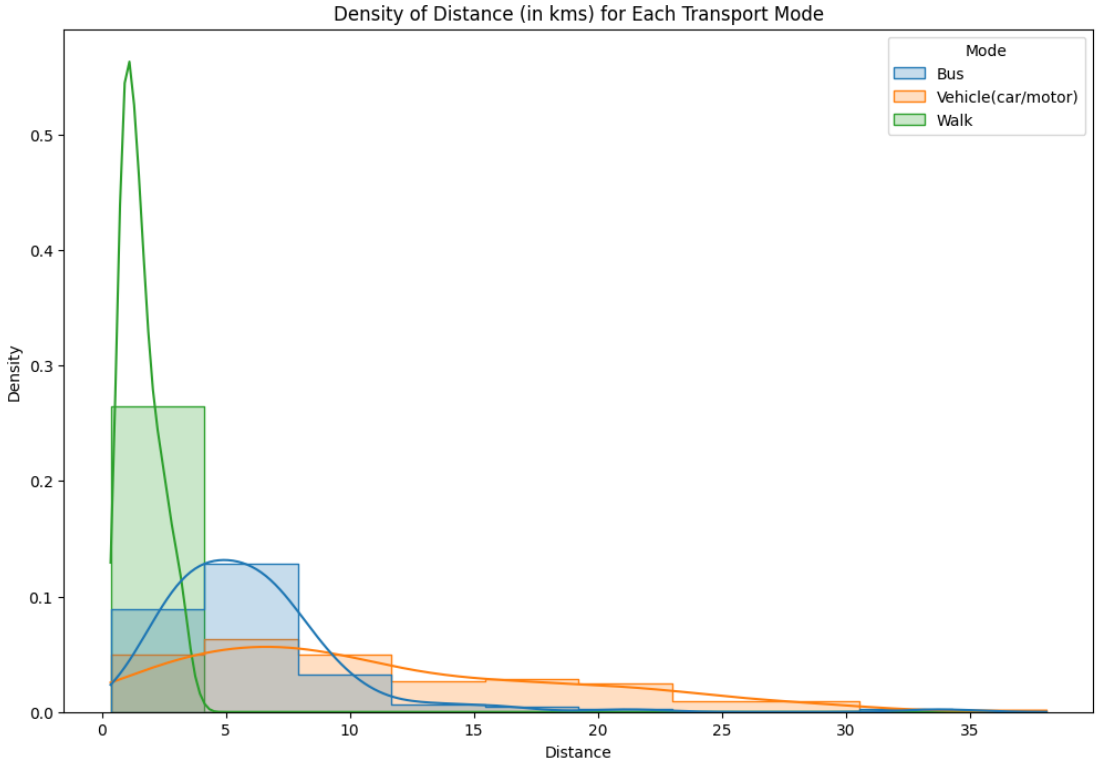


Figure 30 Density of Distance by mode

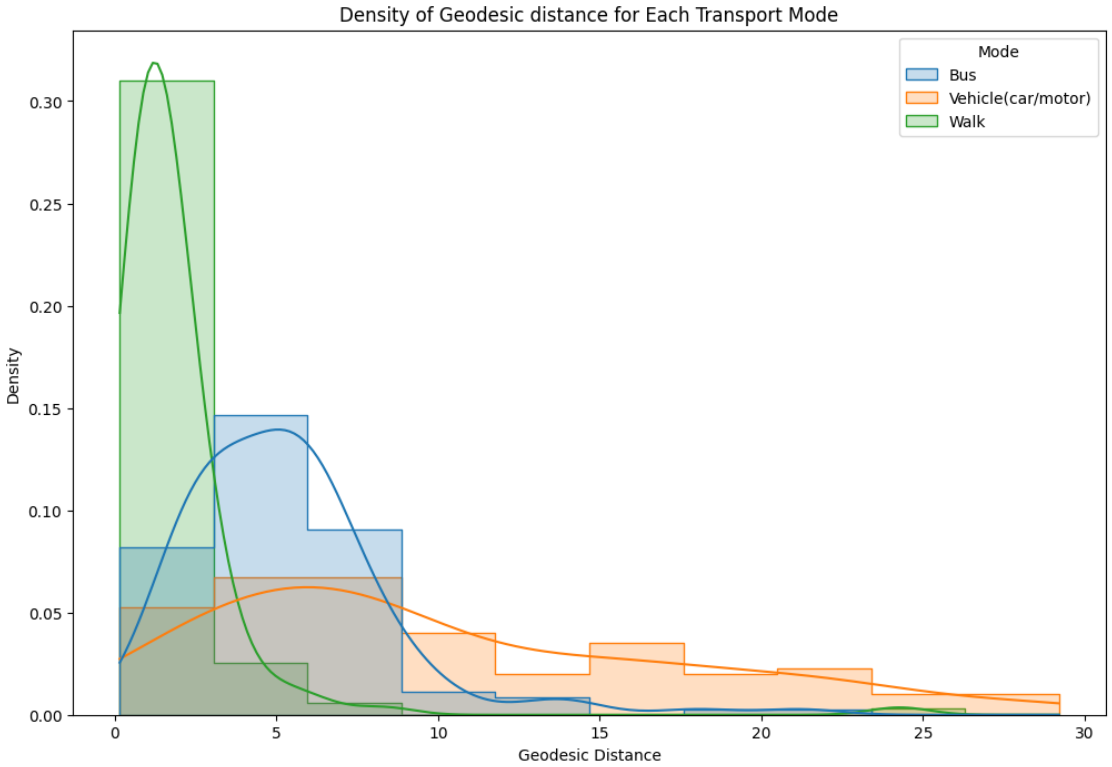


Figure 31. Density of Geodesic Distance by mode

The final feature concerns the time required (in minutes) for respondents to commute to work, measured in minutes. It can be observed that it follows a distribution like the distance features, indicating a positive skew, with many commutes occurring in the initial classes of approximately 3 to 45 minutes. The histogram, along with the descriptive statistics, are depicted below.

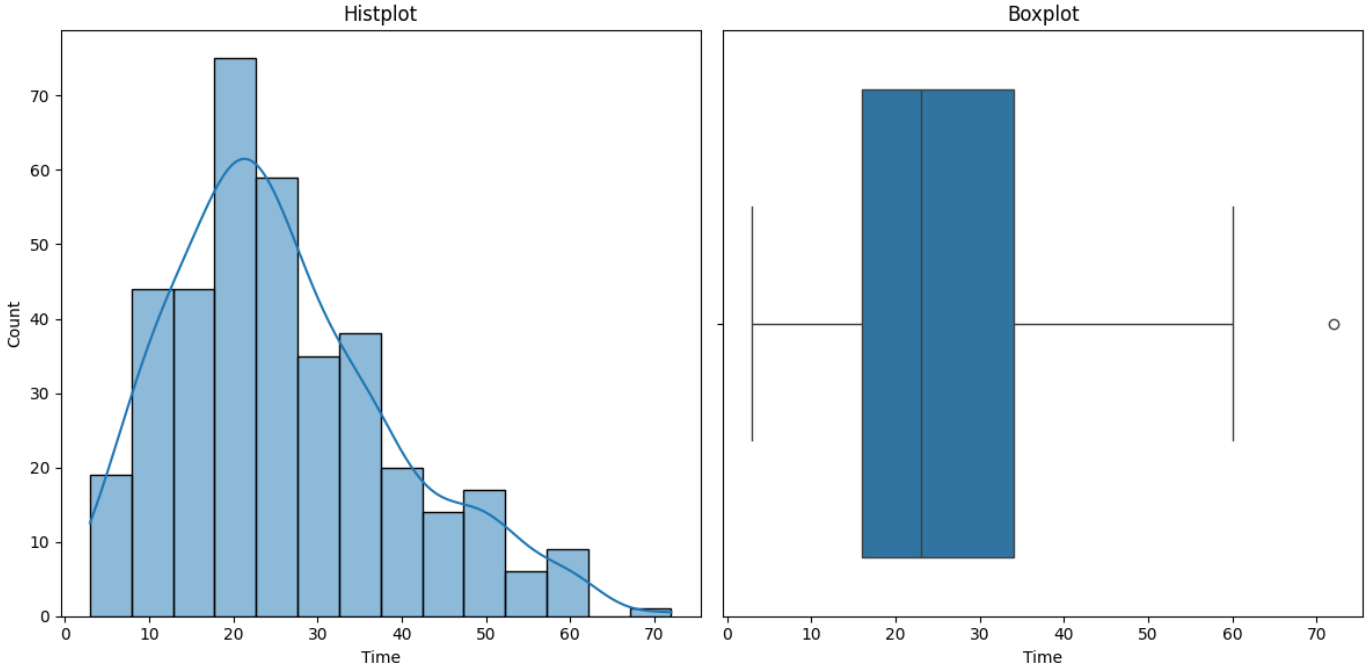
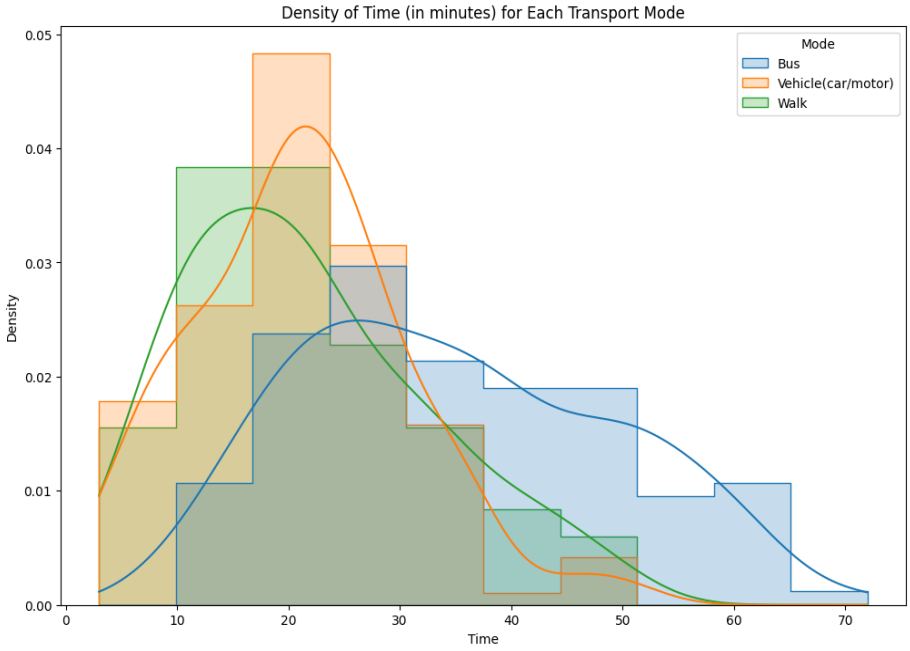


Figure 32. Duration (in minutes)



	Time
Mean	25.78
std	13.3
min	3
25 %	16
50 %	23
75 %	34
max	72

Figure 33. Density of duration and descriptives

The density plot was again constructed for time and each transport mode. For bus density is more spread across the X-axis while walking and private vehicles are more concentrated within 5- and 40-minutes commute time.

Additional graphs were also generated, to better understand the data. The following graph illustrates the departure time grouped by transport mode. At "peak" hour, typically between 06.00 and 09.00 most of the observations are for private vehicles. Though, as departure time progresses bus and walk commuters overcome those who commute with private vehicles, specifically during 12.00 and 15.00.

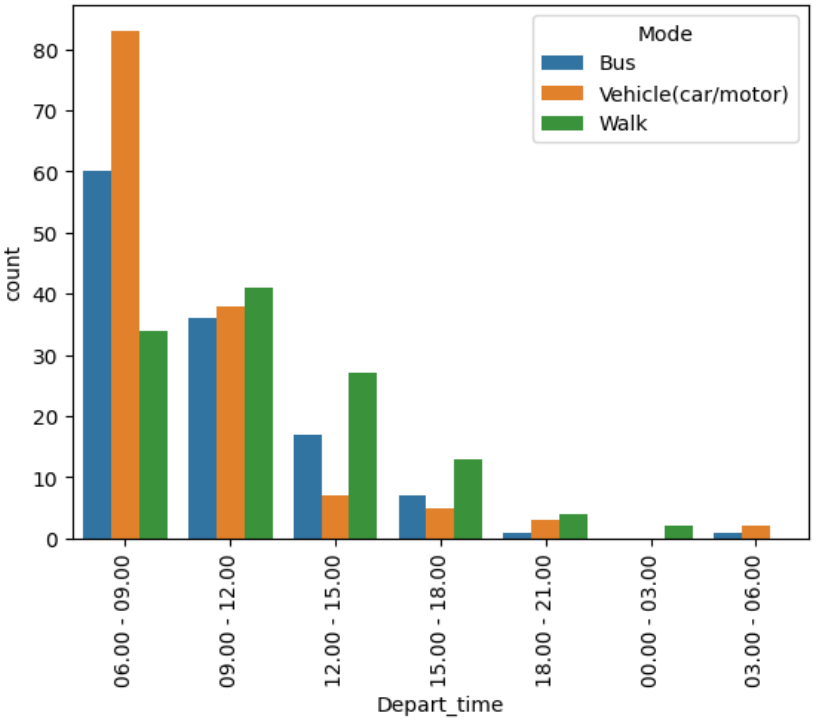


Figure 34. Departure time by mode

The subsequent graph displays the age of the respondents categorized by transport mode. It is apparent that most respondents below 20 years old refrain from commuting to work using a private vehicle. In contrast, as age increases, there is a tendency for individuals to commute to work using a private vehicle. Fewer people opt for bus or walking as their mode of commuting beyond the age of 41.

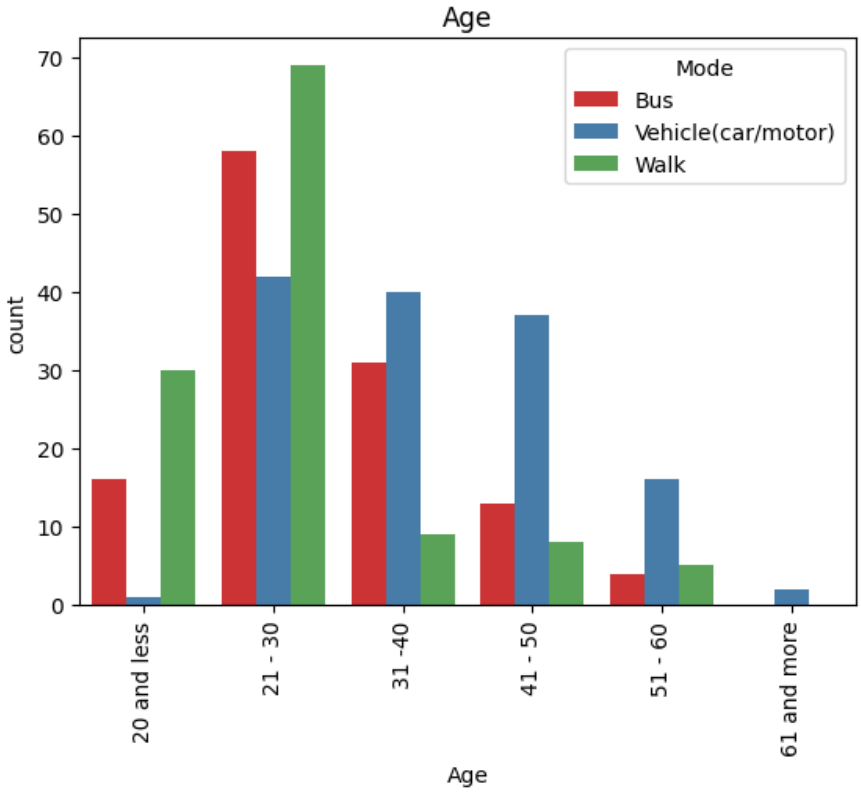


Figure 35. Age by mode

The following graph illustrates the income range grouped by transport mode. The trend is likewise the age of the respondent. When income falls within the 0 to 500 range, commuters predominantly opt for either bus or walking. Conversely, as income increases, private vehicles become the apparent choice for commuting.

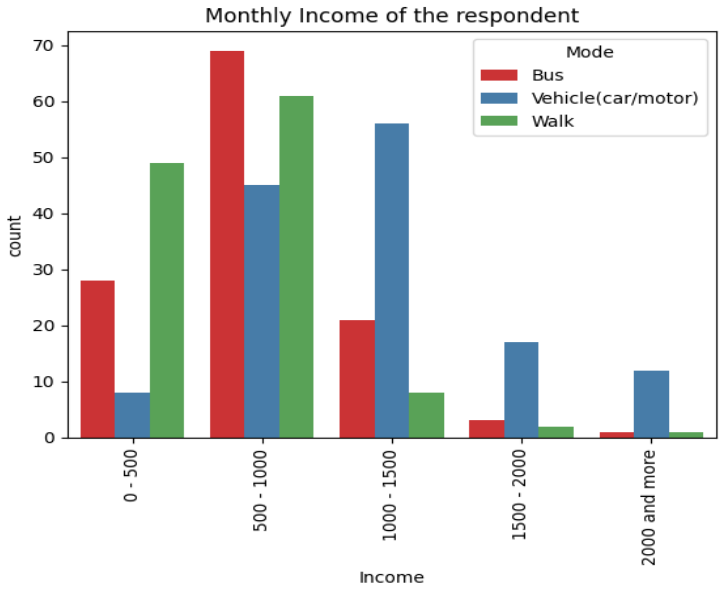


Figure 36. Income by mode

4.1.3. Data Preprocess

To prepare the data for the classification model, a preprocessing step is necessary to ensure the models function effectively. It is crucial to encode all features into numerical form, as models can only process numerical data. Firstly, low frequencies in some categories were grouped so to not lose information from the dataset. More specifically for Age and Depart time low frequencies were grouped as depicted below.

```
perc=df['Age'].value_counts(True)*100
print(perc)
```

```
Age
21 - 30        44.356955
31 -40         20.997375
41 - 50        15.223097
20 and less    12.335958
51 - 60         6.561680
61 and more     0.524934
Name: proportion, dtype: float64
```

```
value_mapping = {"51 - 60":'51 and more', "61 and more":'51 and more'}
df['Age'] = df['Age'].replace(value_mapping)
```

```
perc=df['Age'].value_counts(True)*100
print(perc)
```

```
Age
21 - 30        44.356955
31 -40         20.997375
41 - 50        15.223097
20 and less    12.335958
51 and more     7.086614
Name: proportion, dtype: float64
```

```
value_mapping = {"15.00 - 18.00":'Other', "18.00 - 21.00":'Other', "03.00 - 06.00":"Other", "00.00 - 03.00":"Other"}
df['Depart_time'] = df['Depart_time'].replace(value_mapping)
```

```
perc=df['Depart_time'].value_counts(True)*100
print(perc)
```

```
Depart_time
06.00 - 09.00    46.456693
09.00 - 12.00    30.183727
12.00 - 15.00    13.385827
Other             9.973753
Name: proportion, dtype: float64
```

For Age “51 – 60” and “more than 61” responses were grouped into one category, “51 and more. For Depart Time responses outside of the 06.00 - 15.00 range were grouped into “Other” category. The Income feature was numerically encoded by selecting the mean of each range as its value. For instance, the value 250 was assigned for the income range "0-500," 750 for "500-1000," and so forth. The encoding for Income is illustrated below.

```
value_mapping = {"0 - 500":250, "500 - 1000":750, "1000 - 1500": 1250, "1500 - 2000":1750, "2000 and more":225
df['Income'] = df['Income'].replace(value_mapping)
```

```
perc=df['Income'].value_counts(True)*100
print(perc)
```

```
Income
750     45.931759
250     22.309711
1250    22.309711
1750     5.774278
2250     3.674541
```

Additionally, the impact factors (cost, convenience, safety, health, environment, parking, weather) were encoded into an ordinal scale of 1-5. More specifically the encoding is illustrated below:

```
value_mapping = {"Totally disagree":1, "Disagree":2, "Neutral":3, "Agree":4, "Totally agree":5}
df['Convenience'] = df['Convenience'].replace(value_mapping)
```

```
value_mapping = {"Totally disagree":1, "Disagree":2, "Neutral":3, "Agree":4, "Totally agree":5}
df['Cost'] = df['Cost'].replace(value_mapping)
```

```
value_mapping = {"Totally disagree":1, "Disagree":2, "Neutral":3, "Agree":4, "Totally agree":5}
df['Safety'] = df['Safety'].replace(value_mapping)
```

```
value_mapping = {"Totally disagree":1, "Disagree":2, "Neutral":3, "Agree":4, "Totally agree":5}
df['Health'] = df['Health'].replace(value_mapping)
```

```
value_mapping = {"Totally disagree":1, "Disagree":2, "Neutral":3, "Agree":4, "Totally agree":5}
df['Weather'] = df['Weather'].replace(value_mapping)
```

```
value_mapping = {"Totally disagree":1, "Disagree":2, "Neutral":3, "Agree":4, "Totally agree":5}
df['Environment'] = df['Environment'].replace(value_mapping)
```

```
value_mapping = {"Totally disagree":1, "Disagree":2, "Neutral":3, "Agree":4, "Totally agree":5}
df['Parking'] = df['Parking'].replace(value_mapping)
```

Furthermore, all the binary features were encoded into 0 and 1. This process essentially assigns 0 for "No" and 1 for "Yes" observations. Gender is also part of the label encoder process with 0 assigned to "Males" and 1 to "Females".

The last features that require encoding are Age and Depart_time. For those features one-hot encoding process was used. This procedure essentially generates new binary features for each potential answer, assigning values of 0 to indicate the absence of that value in the specific row and 1 to imply the presence of that value in the row. The process is illustrated below.

```
one_hot_encoded = pd.get_dummies(df[['Age', 'Depart_time']])
one_hot_encoded = one_hot_encoded.astype(int)
df = pd.concat([df, one_hot_encoded], axis=1)
```

An example of the results from the one hot encoding process is depicted below.

Age_20 and less	Age_21 - 30	Age_31 -40	Age_41 - 50	Age_51 and more	Depart_time_06.00 - 09.00	Depart_time_09.00 - 12.00
0	1	0	0	0	0	1
0	1	0	0	0	1	0
0	1	0	0	0	1	0
0	1	0	0	0	1	0
0	1	0	0	0	1	0

The next step was to display the correlation matrix. It is crucial to examine correlations among various features to determine which ones to retain and which ones to drop prior to the modelling application. These decisions rely on the correlation coefficient values, which can range from -1 to 1. Values of -1 and 1 indicate a perfect correlation (positive or negative) between variables, potentially leading to multicollinearity issues that could hinder the performance of the models. Hence, highly correlated features should be eliminated before the application process.

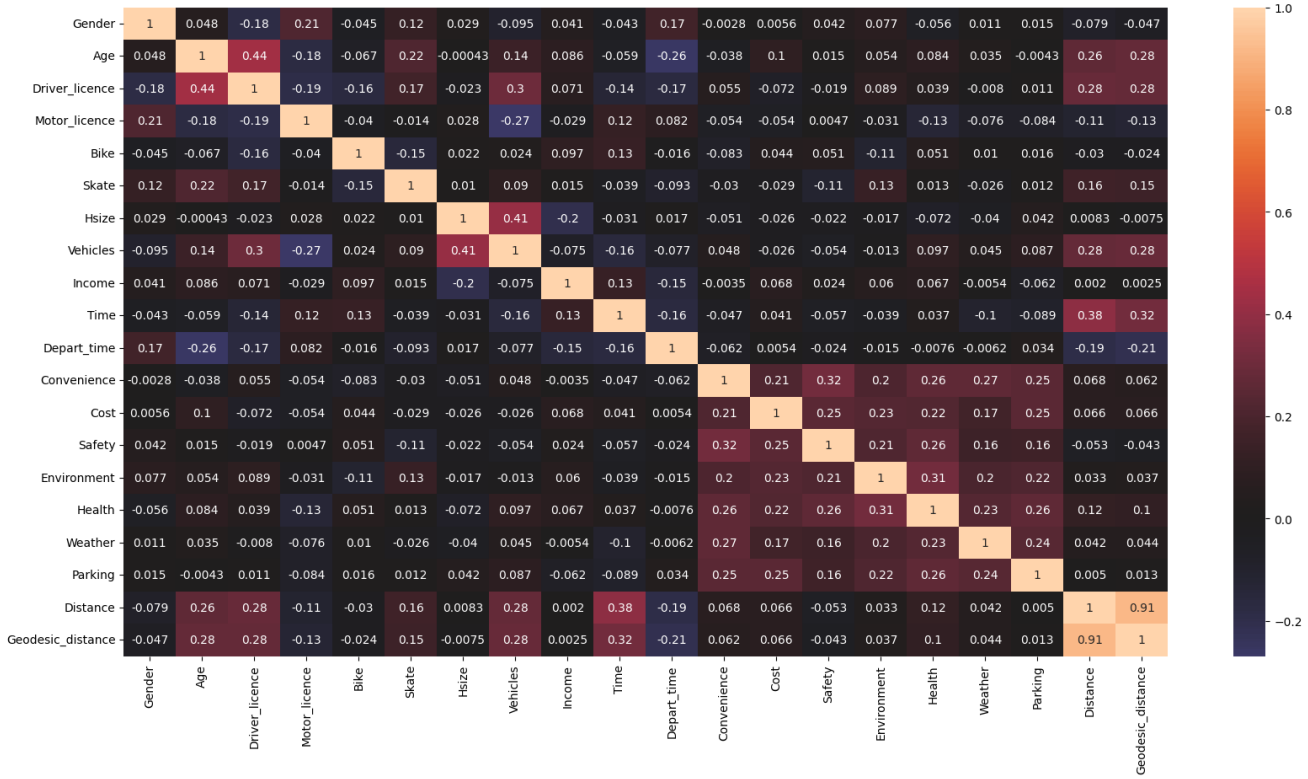


Figure 37. Correlation matrix

From the figure above, it is identified that there are strong correlations in a few cases. Distance and Geodesic_distance with a value of 0.91, followed by Driver_licence and Age with 0.44. Only one the Distance metrics will be used for training the models each time. The "vehicles" feature will also be excluded because, when it takes a value of 0 (indicating no household vehicles), it introduces a bias in the classifier, making it easier to classify non-private vehicle observations. Skate, Driver licence, Bike and Motor licence features will also be excluded from the models for the same reason. Additionally, to review multicollinearity issues the VIF (Variance Inflator Factor) function was used. VIF measures the severity of multicollinearity, so each feature is assigned a value. Values above 10 raise concerns about the correlation of the features and should be removed. The results of the first attempt using VIF from statsmodels library are depicted below.

	Variable	VIF
0	const	0.000000
1	Gender	1.129359
2	Hsize	1.049693
3	Income	1.681202
4	Time	1.557271
5	Convenience	1.508518
6	Cost	1.198751
7	Safety	1.883755
8	Environment	1.513632
9	Health	1.732697
10	Weather	1.446991
11	Parking	1.470916
12	Distance	1.490938
13	Age_20 and less	inf
14	Age_21 - 30	inf
15	Age_31 -40	inf
16	Age_41 - 50	inf
17	Age_51 and more	inf
18	Depart_time_06.00 - 09.00	inf
19	Depart_time_09.00 - 12.00	inf
20	Depart_time_12.00 - 15.00	inf
21	Depart_time_Other	inf

Figure 38. VIF 1st try

The occurrence of "inf" values in certain features points to multicollinearity issues, particularly in one hot encoded features. This arises because, during the one-hot encoding process, one of the resulting binary features must be omitted to prevent multicollinearity. Consequently, depart time 06:00-00, and Age 21-30 were excluded. The VIF function was subsequently reapplied, yielding the following values.

	Variable	VIF
0	const	63.846924
1	Gender	1.129359
2	Hsize	1.049693
3	Income	1.681202
4	Time	1.557271
5	Convenience	1.508518
6	Cost	1.198751
7	Safety	1.883755
8	Environment	1.513632
9	Health	1.732697
10	Weather	1.446991
11	Parking	1.470916
12	Distance	1.490938
13	Age_20 and less	1.261247
14	Age_31 -40	1.432278
15	Age_41 - 50	1.420269
16	Age_51 and more	1.303088
17	Depart_time_09.00 - 12.00	1.322222
18	Depart_time_12.00 - 15.00	1.352682
19	Depart_time_Other	1.273932

Figure 39. VIF 2nd try

Now, with all VIF values below 5, these finalized features are set to be utilized in the models. The dataset was initially shuffled and consequently split into training and testing sets. The train_test_split() function of sklean was used with a ratio of 60:40. Hence, 60% of the data (228 samples) will train the models, while 40% (153) will be used for evaluation. The data was subsequently normalized using the MinMaxScaler() function from sklearn, which transforms the data values into a 0 – 1 range. The dataframes before and after the normalization process are depicted below.

Income	Time	Convenience	Cost	Safety	Environment	Health	Weather	Parking	Distance	Age_20 and less
750	26	4	3	4	4	4	3	5	21.8	0
2250	13	5	1	5	5	3	5	5	4.0	0
750	31	3	4	3	3	3	3	2	4.6	0
250	19	3	3	3	4	4	4	3	1.3	1
750	6	4	5	5	5	5	5	1	0.8	0

Income	Time	Convenience	Cost	Safety	Environment	Health	Weather	Parking	Distance	Age_20 and less
0.25	0.403509	0.75	0.50	0.75	0.75	0.75	0.50	1.00	0.568212	0.0
1.00	0.175439	1.00	0.00	1.00	1.00	0.50	1.00	1.00	0.096689	0.0
0.25	0.491228	0.50	0.75	0.50	0.50	0.50	0.50	0.25	0.112583	0.0
0.00	0.280702	0.50	0.50	0.50	0.75	0.75	0.75	0.50	0.025166	1.0
0.25	0.052632	0.75	1.00	1.00	1.00	1.00	1.00	0.00	0.011921	0.0

Figure 40. Normalization

4.1.4. Decision Tree

The first model that was applied was that of the decision tree. The DecisionTreeClassifier(random_state = 21) function was used to create the model. The random_state parameter controls the randomness and ensures reproducibility in the results. The model was fitted on the training set using the fit() command and consequently evaluated on the test set using the predict() command. The results are illustrated below.

Table 3. Decision tree performance with default parameters

	Precision	Recall	F1	Accuracy
Private Vehicle	0.73	0.78	0.75	0.78
Bus	0.83	0.69	0.76	
Walk	0.79	0.86	0.82	
Macro Average	0.78	0.78	0.78	

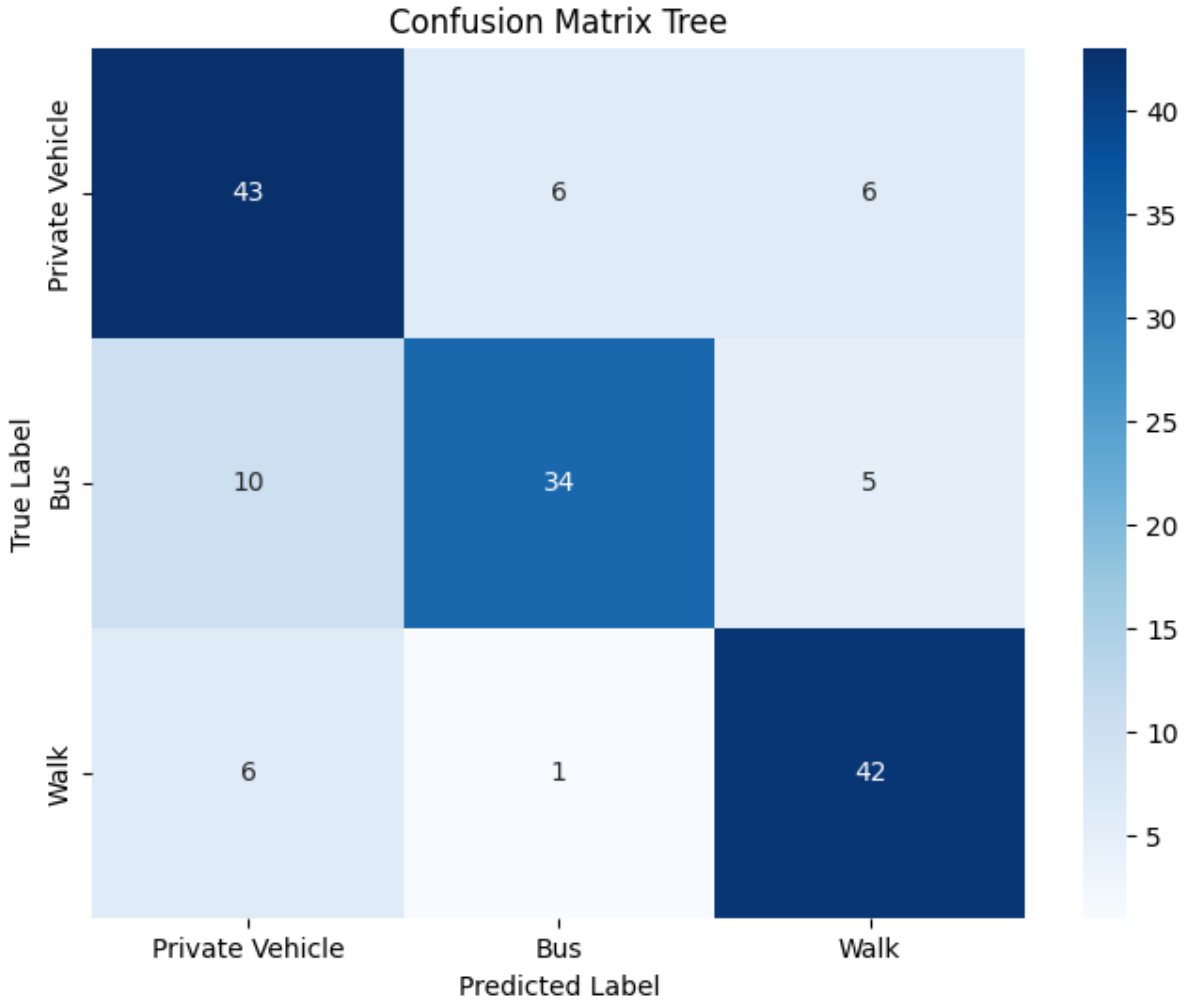


Figure 41. Confusion Matrix - Default Tree

The overall accuracy of the model is 0.78 while the max_depth of the tree was found to be 6. The max depth parameter controls the maximum possible length of the path from the root node to a leaf node. The model correctly predicted 42 out of 49 (0.86) instances as “walk” when the true label was “walk”. For classes “private vehicle” and “bus” recall was found 0.78 and 0.69 respectively. The highest precision is for class “bus” (0.83) indicating that of the 41 predictions the model made as “bus” 34 belonged in that class. The biggest misclassification occurs between private vehicle - bus, where the model predicted 10 instances as “private vehicle” while the true label was “bus”. The model clearly overfits the data since accuracy for the training set is 1, while the performance on the test set was average. To address this concern, a 10-fold cross-validation was implemented to assess whether tree pruning, which involves reducing the size of the tree, enhances the model's performance. The results of cross validation are presented below.

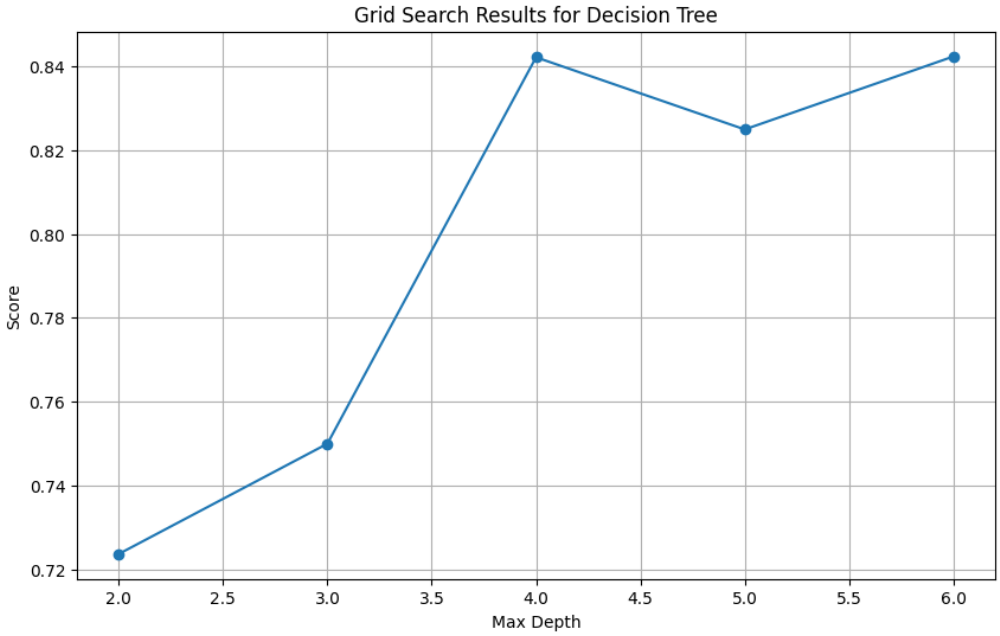


Figure 42. Cross Validation for tree depth

From the validation results it is identified that accuracy is higher for depth equal to 4 and 6. The tree was pruned at max depth value of 4 and was revaluated on the test set. The results are illustrated on the classification report below.

Table 4. Pruned Tree performance

	Precision	Recall	F1	Accuracy
Private Vehicle	0.84	0.76	0.80	0.82
Bus	0.89	0.82	0.85	
Walk	0.76	0.90	0.82	
Macro Average	0.83	0.83	0.82	

Compared to the original tree model, the overall accuracy has been increased from 0.78 to 0.82. Recalls for both bus and walk increased. More particularly, for bus recall increased from 0.69 to 0.82, while for walk it increased from 0.86 to 0.90. In contrast, recall for private vehicle decreased from 0.78 to 0.76 with a trade off in precision as it increased from 0.73 to 0.84. Overall reducing the max depth parameter slightly increased the performance, though training accuracy remains high at 0.95. Hence, additional parameters were tuned for the model. More specifically, min_samples_split and min_samples_leaf parameters were also tuned via GridSearchCV() function. Min_samples_split refers to the number of samples required to split an internal node. In contrast, min_samples_split refers to the number of samples required to be in a leaf mode. The set of parameters for the validation process are depicted below:

```
dt_model = DecisionTreeClassifier(random_state=21)
param_grid = {
    'max_depth': [2, 3, 4, 5],
    'min_samples_split': [1, 2, 3, 4,  5, 6, 7, 8, 9, 10],
    'min_samples_leaf': [1, 2, 3, 4, 5, 6, 7],
}
```

The validation process wielded the following results:

max_depth = 5, **min_samples_split = 2**, **min_samples_leaf = 3 and a validation accuracy of 0.85**. The best model was then retrieved and evaluated on the test set. The results are depicted below.

Table 5. Tuned Tree performance

	Precision	Recall	F1	Accuracy
Private Vehicle	0.82	0.82	0.82	0.83
Bus	0.89	0.80	0.84	
Walk	0.80	0.88	0.83	
Macro Average	0.83	0.83	0.83	

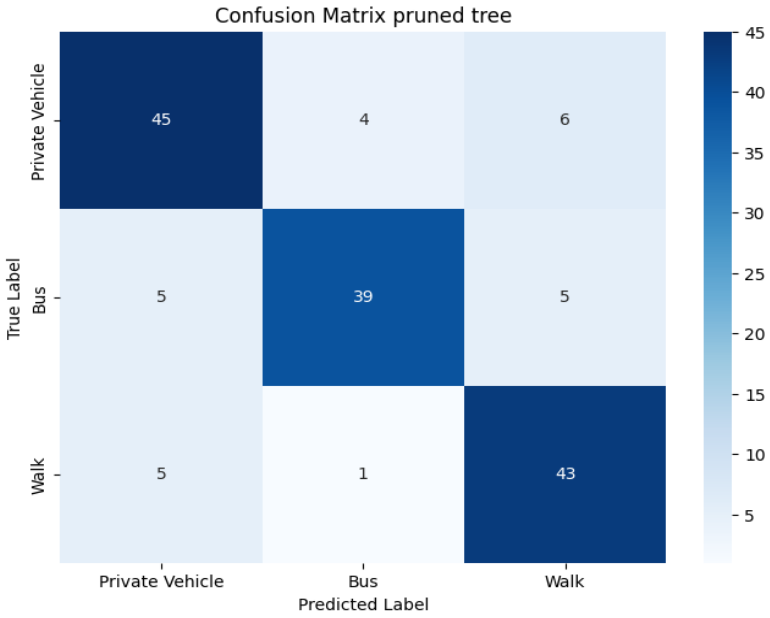


Figure 43. Confusion Matrix - Tuned Tree

The overall accuracy of the decision tree increased from 0.82 to 0.83. Simultaneously, the training accuracy remained the same at 0.95. Compared to the previous tree, recall for "bus" decreased from 0.82 to 0.80, while for private vehicle increased from 0.76 to 0.82. Furthermore, the precision for "walk" also increased from 0.76 to 0.80. The most significant misclassification occurred for walk and private vehicle labels, where the model predicted 6 instances as "walk" but the true label was "private vehicle". Below the tree structure of the pruned tree is illustrated.

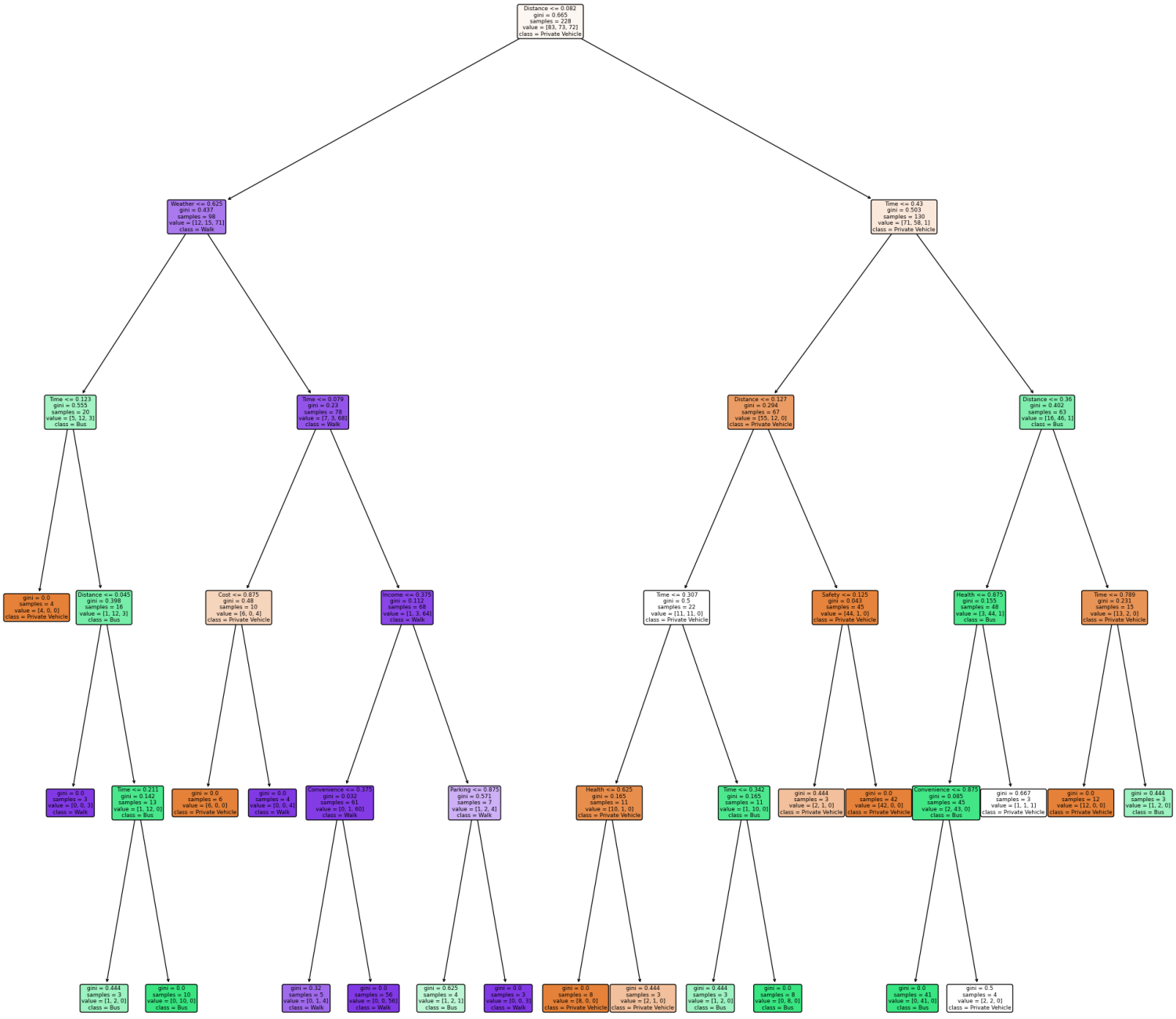


Figure 44. Tuned Tree structure

Finally, the feature importances of the model were retrieved using the feature_importances_ function. Results are illustrated below:

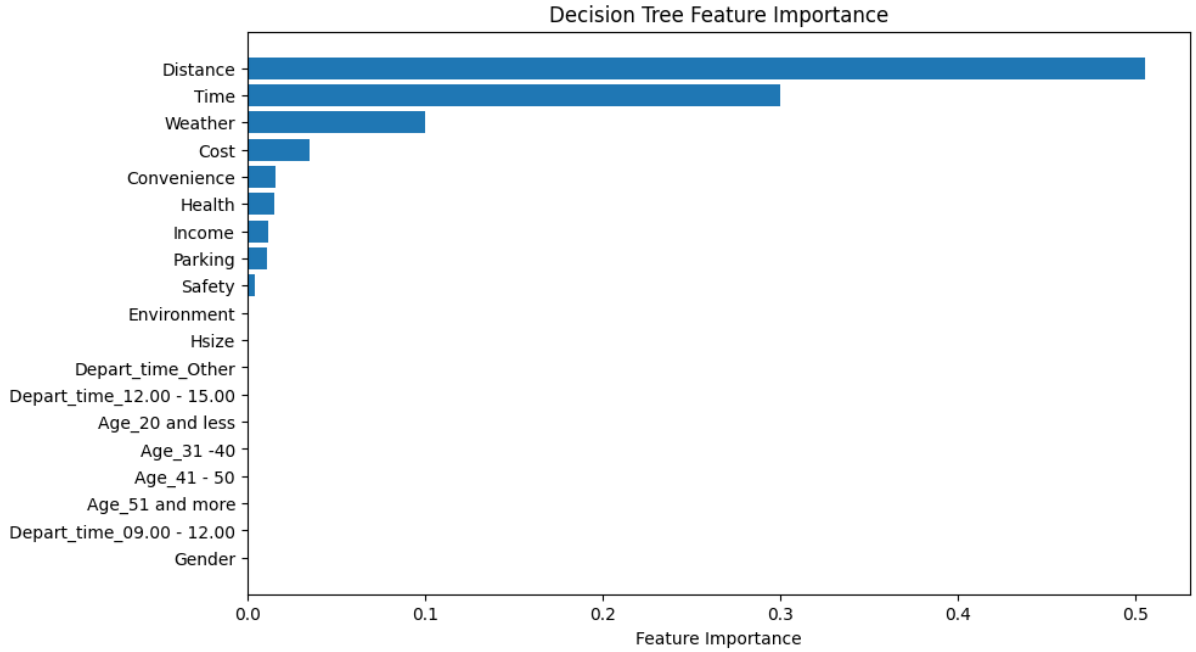


Figure 45. Decision Tree feature importance

The feature importance essentially shows how much each feature contributed to the splits that the tree made. The plot above reveals that Distance and Time features are identified as the most crucial factors for the models, followed by Weather and Cost. It is noteworthy that only 9 features out of the 19 available were selected to construct the decision tree.

The next experimentation was using the Geodesic_distance instead of the Distance feature. The model was trained on the training data and evaluated on the test set. The results are illustrated below.

Table 6. Decision Tree - Geodesic Distance

	Precision	Recall	F1	Accuracy
Private Vehicle	0.82	0.89	0.85	0.82
Bus	0.78	0.82	0.80	
Walk	0.86	0.73	0.79	
Macro Average	0.82	0.81	0.81	

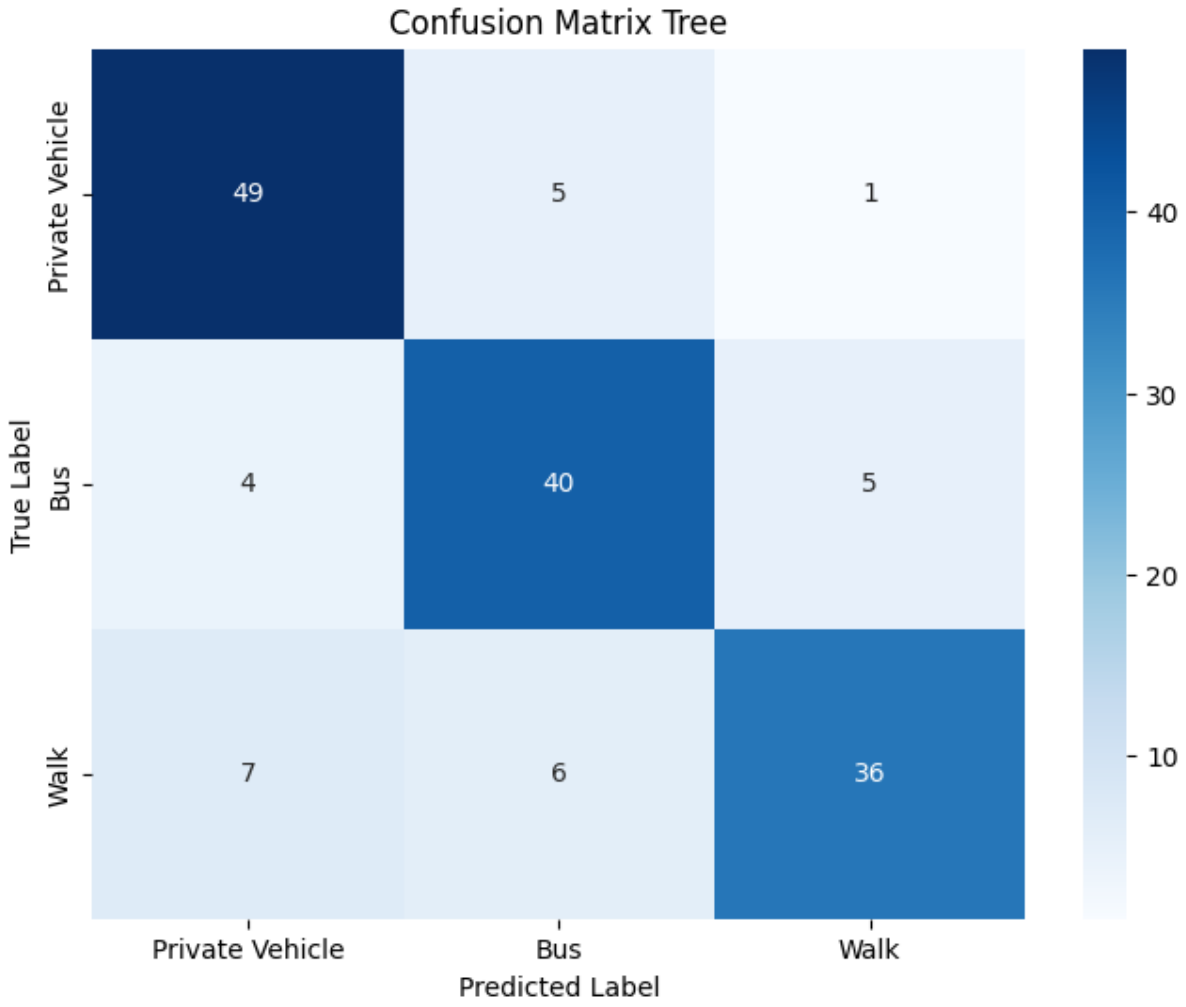


Figure 46. Confusion Matrix - Tree - Geodesic Distance

The accuracy of the model using the geodesic_distance was at 0.82. Recall for private vehicle and bus were at 0.89 and 0.82 respectively. Walk has the lowest recall for 0.73, with a trade off in precision which is the highest with 0.86. The highest misclassification is between private vehicle and walk as the model predicted 7 instances as private vehicle while the true label was walk. Again, training accuracy is at 1 so parameter tuning was explored to identify if the performance would increase. The tree was again pruned using cross validation. The parameters explored are illustrated below.

```
dt_model = DecisionTreeClassifier(random_state=21)
param_grid = {
    'max_depth': [2, 3, 4, 5, 6, 7, 8],
    'min_samples_split': [1, 2, 3, 4,  5, 6, 7, 8, 9, 10],
    'min_samples_leaf': [1, 2, 3, 4, 5, 6, 7],
}
```

The best parameters were **max depth = 6**, **min samples leaf =2** and **min samples split =2** with a **validation accuracy of 0.837**. The best model was then retrieved and evaluated on the test set. The results are depicted below.

Table 7. Pruned Tree - Geodesic Distance

	Precision	Recall	F1	Accuracy
Private Vehicle	0.81	0.87	0.84	0.82
Bus	0.78	0.82	0.80	
Walk	0.88	0.78	0.83	
Macro Average	0.83	0.82	0.82	

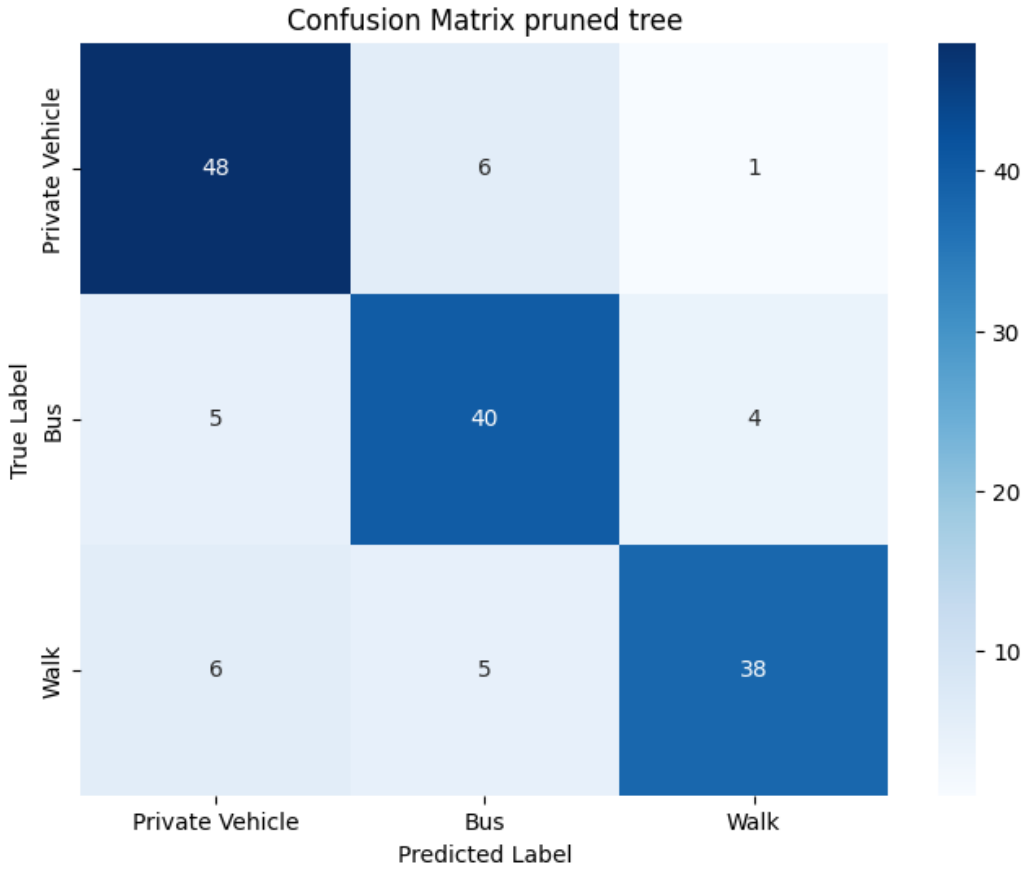


Figure 47. Confusion Matrix - Pruned Tree - Geodesic Distance

The overall accuracy of the model remained the same at 0.82 while training accuracy dropped from 1 to 0.95. Recall for walk increased from 0.73 to 0.78 while for private vehicle in dropped from 0.89 to 0.87. Precision also increased for walk from 0.86 to 0.88. Both Recall and Precision remained the same for Bus. Six instances were misclassified as private vehicle and bus while the true labels were walk and private vehicle respectively. The tree structure using Geodesic_distance is depicted below.

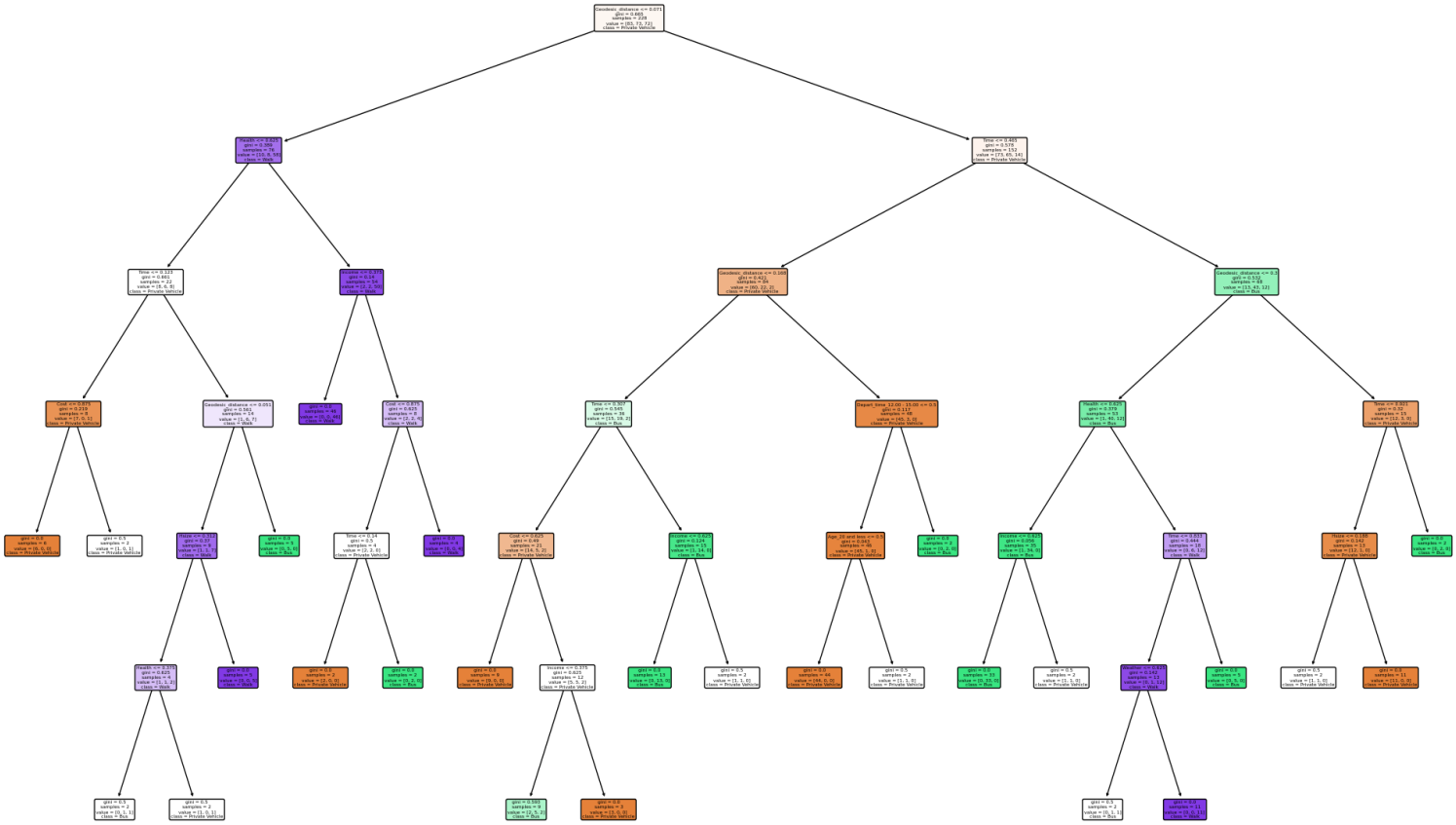


Figure 48. Tree structure - Geodesic Distance

Finally, the feature importances were extracted for the tree using the geodesic distance. The results are depicted below.

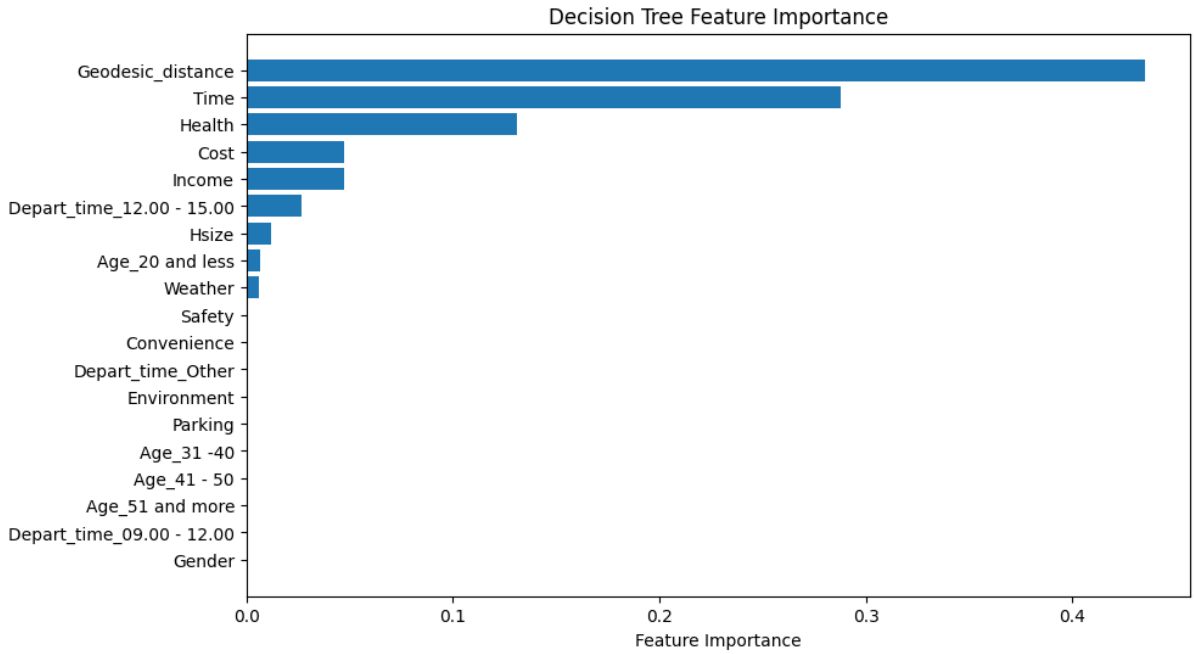


Figure 49. Tree feature importance - Geodesic Distance

Following the trend of the previous model, geodesic distance and time are the most important features for constructing the tree, followed by health cost and income. The model used 9 features in total, out of 19 for constructing the tree like the previous tree utilizing Distance instead.
The following table illustrates the metrics for the two models using different distance metrics.

Table 8. Pruned Tree evaluation (Distance vs Geodesic Distance)

	Pruned Tree (Distance)			Pruned Tree (Geodesic distance)		
	Private vehicle	Bus	Walk	Private vehicle	Bus	Walk
Precision	0.82	0.89	0.80	0.81	0.78	0.88
Recall	0.82	0.80	0.88	0.87	0.82	0.78
F1-score	0.82	0.84	0.83	0.84	0.80	0.83
Accuracy	0.83			0.82		
AUC-Macro	0.90			0.88		

Swapping geodesic distance significantly dropped recall for walk labels, from 0.88 to 0.78. In contrast, there is a slight increase for private vehicle and bus in terms of recall. The overall accuracy using geodesic distance is slightly lower 0.82 compared to 0.83. The auc macro score, representing the average area under curve score for each class, is also slightly higher using distance instead of geodesic. The first model demonstrates a more balanced performance overall.

4.1.5. Random Forest

The next model that was implemented was that of Random Forest. The RandomForestClassifier() function of sklearn was used to create the model. The model was fitted on the training data, followed by evaluation on the test set. The results are illustrated below:

Table 9. Random Forest performance - Default

	Precision	Recall	F1	Accuracy
Private Vehicle	0.87	0.95	0.90	0.88
Bus	0.95	0.78	0.85	
Walk	0.85	0.92	0.88	
Macro Average	0.89	0.88	0.88	

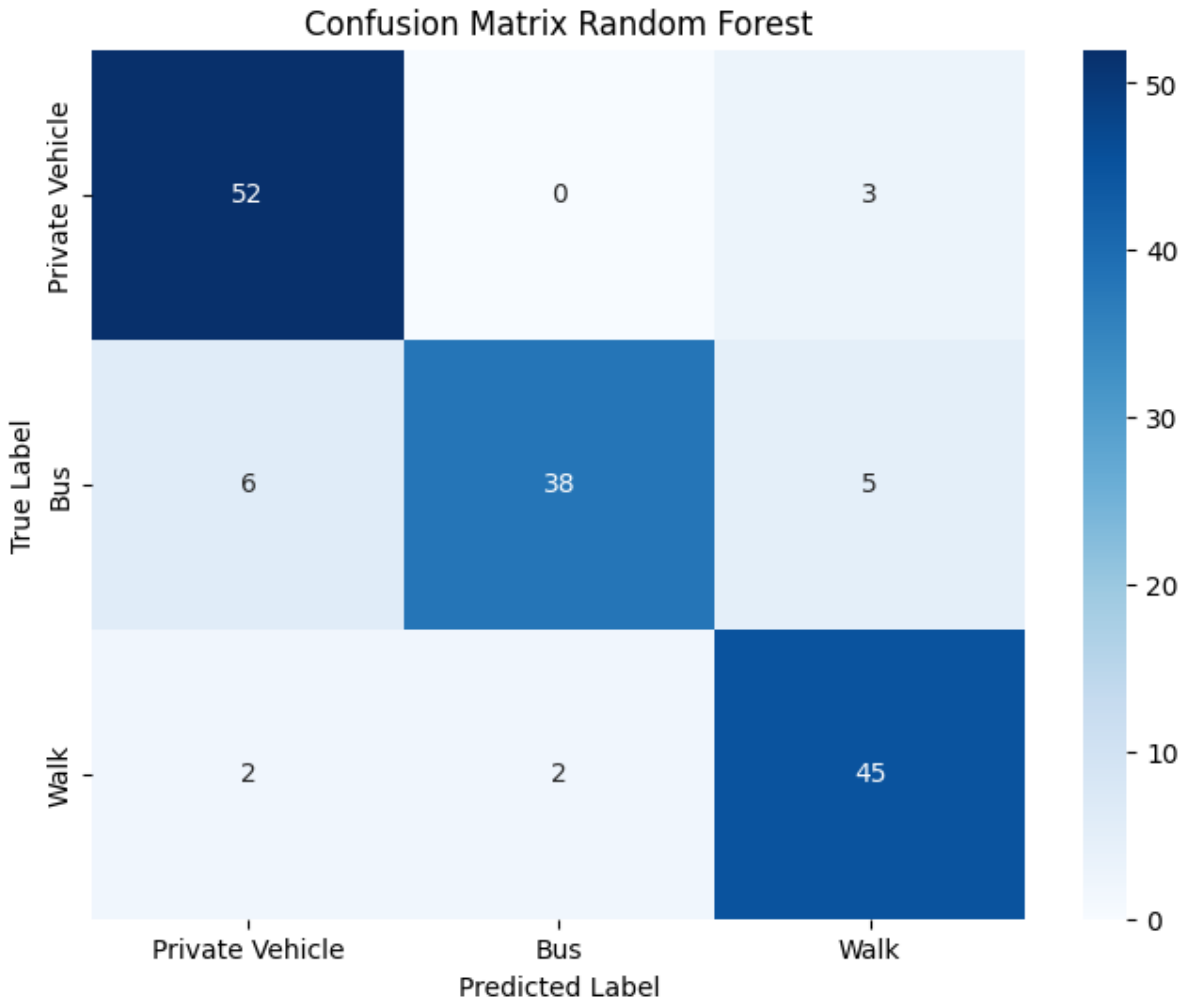


Figure 50. Confusion Matrix - Random Forest - Default

The out of bag error, representing the average error for each predicted outcome and calculated using predictions from the trees that do not include that specific data point in their respective bootstrap sample, was 0.1053. The overall accuracy of the basic model is 0.88, correctly predicting 135 out of 153 total instances. Additionally, the model accurately predicted 38 out of 49 (0.78) cases as "bus" when the true label was "bus." Recall for private vehicle and walk was 0.95 and 0.92 respectively. The highest precision is for bus mode accurately predicting 38 out of 40 predictions made as bus. The highest misclassification is between bus and private vehicle labels as the model predicted 6 instances as private vehicle while the true label was bus. The default values of the model are max_features = "auto" and n_estimators = 100 and max_depth='None'. The max_features parameter refers to the number of features considered each time a split occurs on the decision tree, while "auto" is the square root of the total number of features. In contrast, the n_estimators refer to the number of trees the model builds. Max_depth, which is the depth of each individual tree, is set to None indicating that each individual tree grows without limit. Similar to the decision tree, the GridSearchCV() function was used to tune the model and find the optimal parameters attempting to increase the performance of the model. The process

was run for each parameter individually to identify optimal area for a final grid search. Since the train set contains 19 total features, max features parameter was set at a range between 2 and 19. The validation results are depicted below.

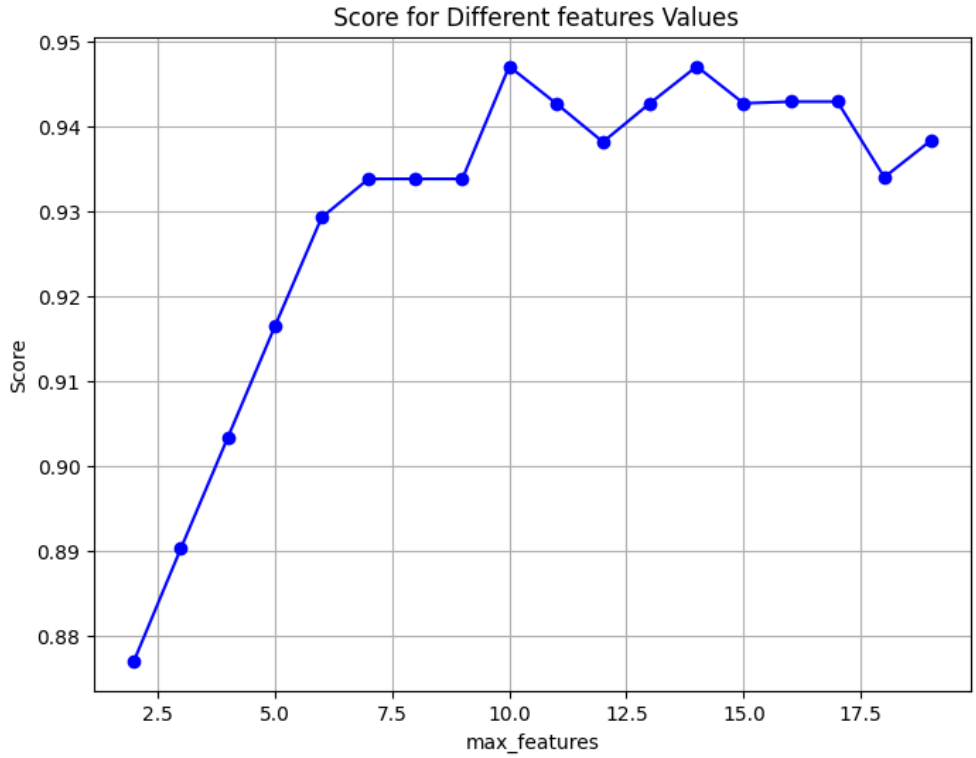


Figure 51. Grid Search RF - Max features

Looking at the plot above, it is evident that the validation accuracy is higher when the number of features is above 6, with the best overall performance to be for 10 number of features and accuracy above 0.94. The next parameter tuned was the number of estimators. The range was set between 50 and 300 number of trees. The validation results are illustrated below.

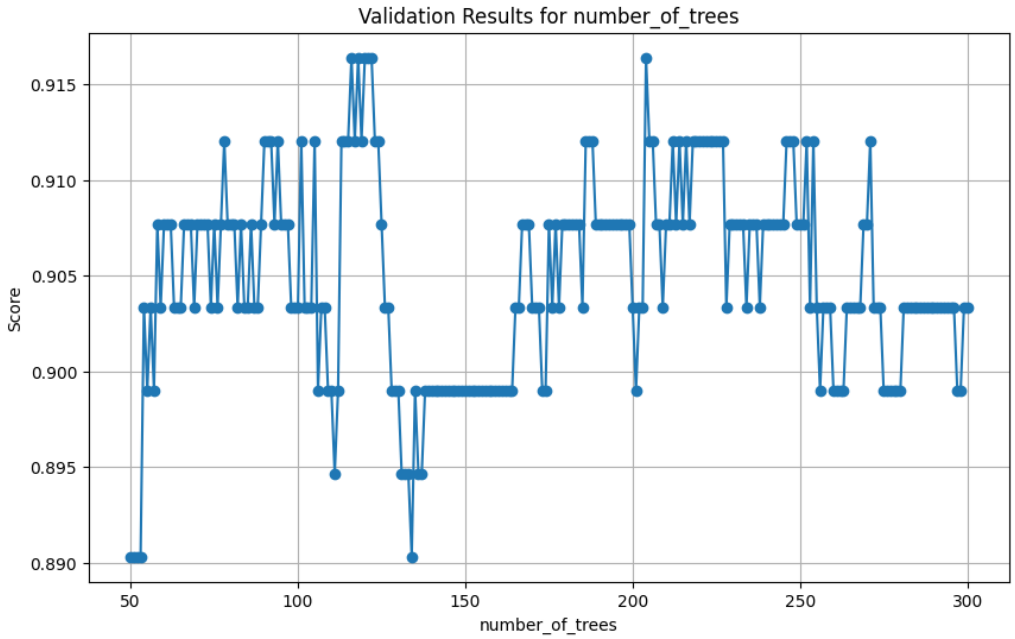


Figure 52. Grid Search RF - Number of estimators

The optimal area, where the validation accuracy is higher, is between 75 and 125 number of trees. The final parameter that was tuned was the max depth of the individual tree. The results are illustrated below.

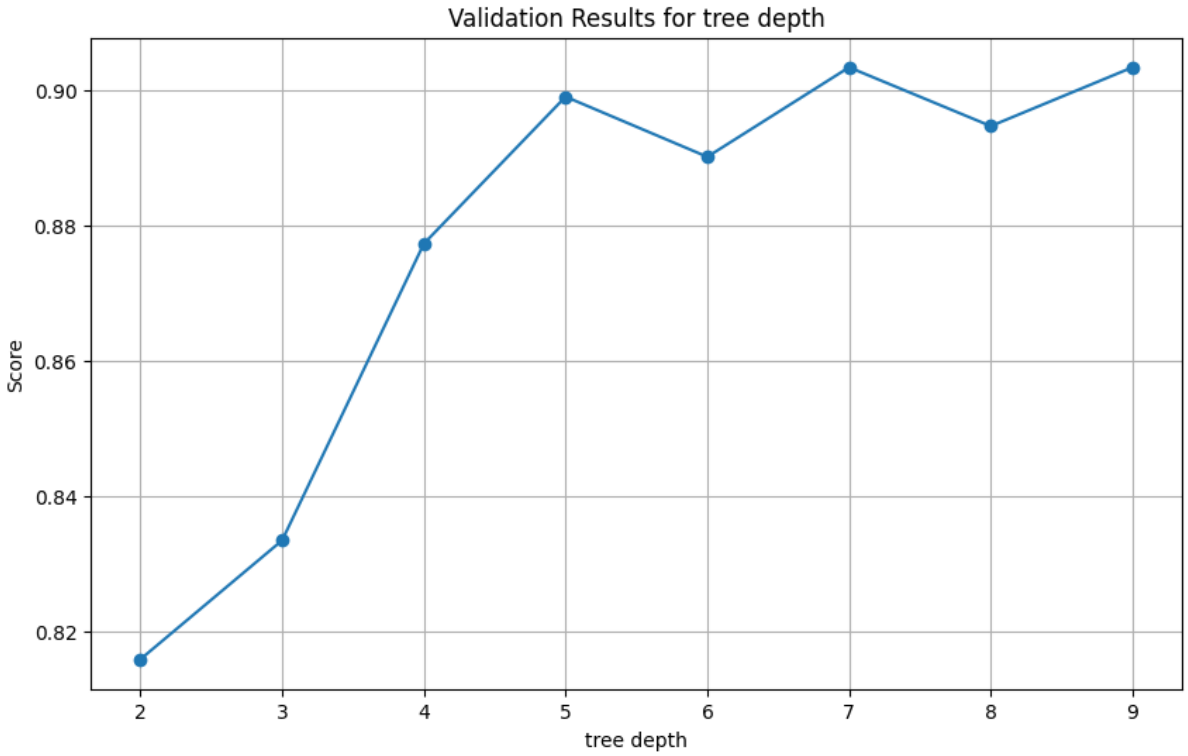


Figure 53. Grid Search RF - Tree depth

The random forest model performs best for a tree depth 7 and 9. After finding the optimal areas, a final grid search was executed with the following parameter values.

```
param_grid = {
    'n_estimators': list(range(75, 126)),
    'max_features': [10, 14],
    'max_depth' : [7, 9]}
```

The results of the final grid search were **max_features = 10** and **n_estimators = 75, and max_depth=7**. Validation accuracy was at 0.95, while the out of bag error was reduced from 0.1053 to 0.0746. The validated model was then retrieved and evaluated on the test set. The results are illustrated below.

Table 10. Tuned Random Forest performance

	Precision	Recall	F1	Accuracy
Private Vehicle	0.96	0.93	0.94	0.92
Bus	0.95	0.86	0.90	
Walk	0.84	0.96	0.90	
Macro Average	0.92	0.91	0.91	

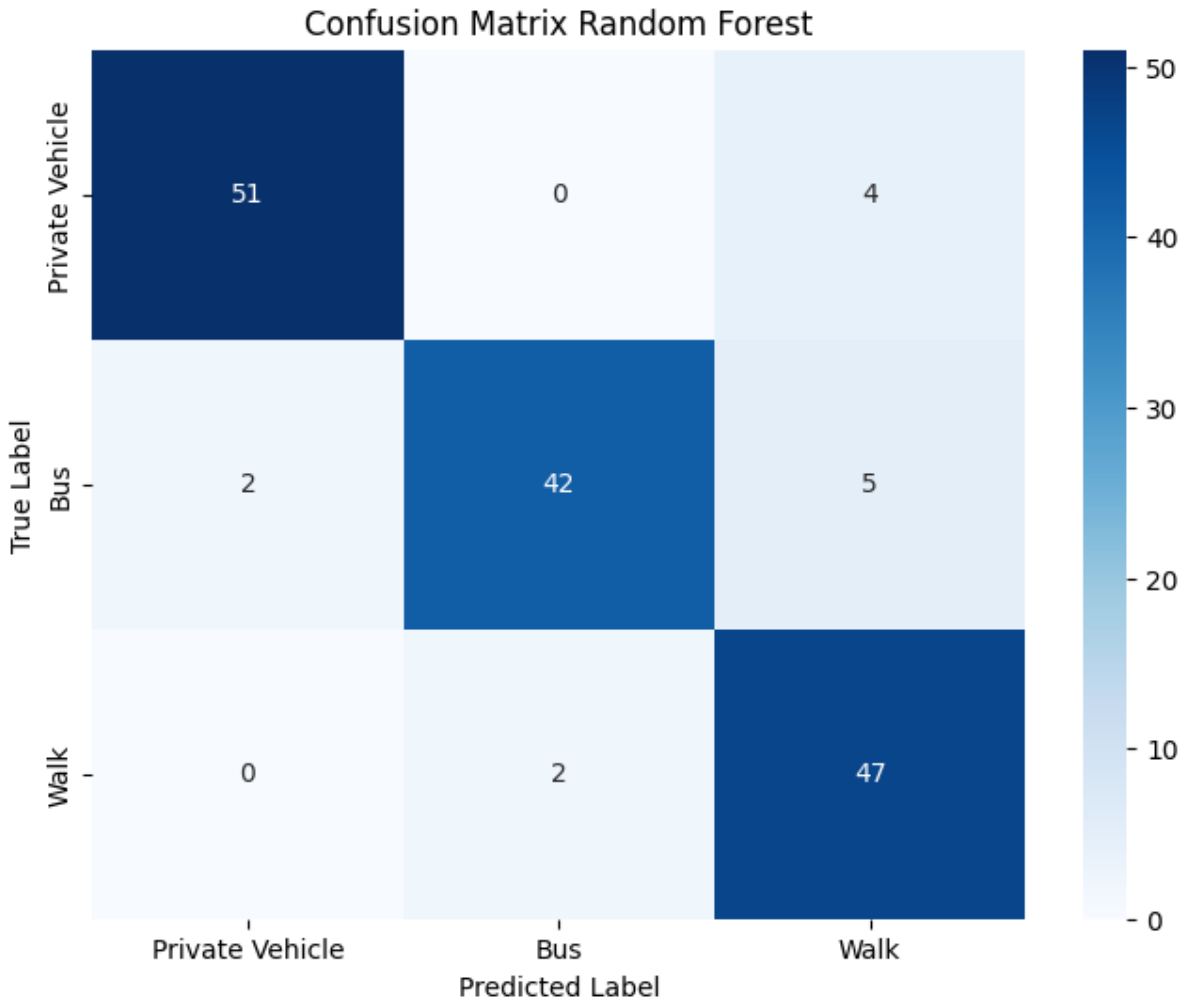


Figure 54. Confusion Matrix - Tuned Random Forest

The overall accuracy of the tuned model increased from 0.88 to 0.92. Recall for bus increased from 0.78 to 0.86 while precision remained at 0.95. Recall was also increased for walk, from 0.92 to 0.96 with a minor drop in precision from 0.85 to 0.84. Precision significantly increased for private vehicle, from 0.87 to 0.96 with a minor decrease in recall from 0.95 to 0.93. The highest misclassification occurs between walk and bus, where the model predicted 5 cases as walk when the true label was bus. The feature importances were also extracted for the model. The results are illustrated below.

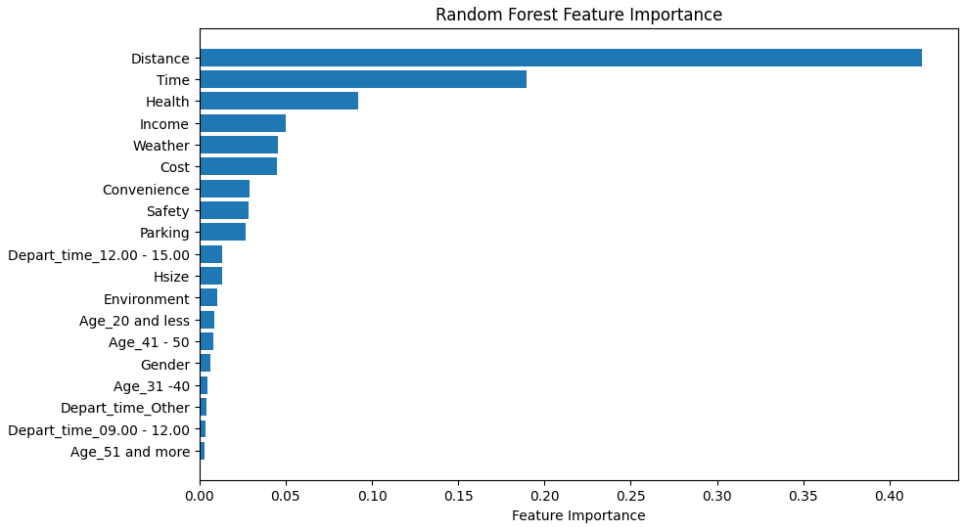


Figure 55. Random Forest feature importance

Like the single decision tree, for random forest feature importance demonstrates how much the feature contributed at the split in each of the individual tree construction. The model prioritises Distance, Time, Health, Convenience, Cost, Safety and Weather as the most crucial features. Unlike the simple tree model, the random forest has essentially utilised all parameters for split criteria at least once.

The last experimentation was swapping Distance with the Geodesic Distance feature. A new model was created, fitted on the training data, and consequently evaluated on the test set. The results are depicted below.

Table 11. Random Forest - Geodesic Distance - Default

	Precision	Recall	F1	Accuracy
Private Vehicle	0.83	0.95	0.88	0.84
Bus	0.88	0.78	0.83	
Walk	0.83	0.80	0.81	
Macro Average	0.85	0.84	0.84	

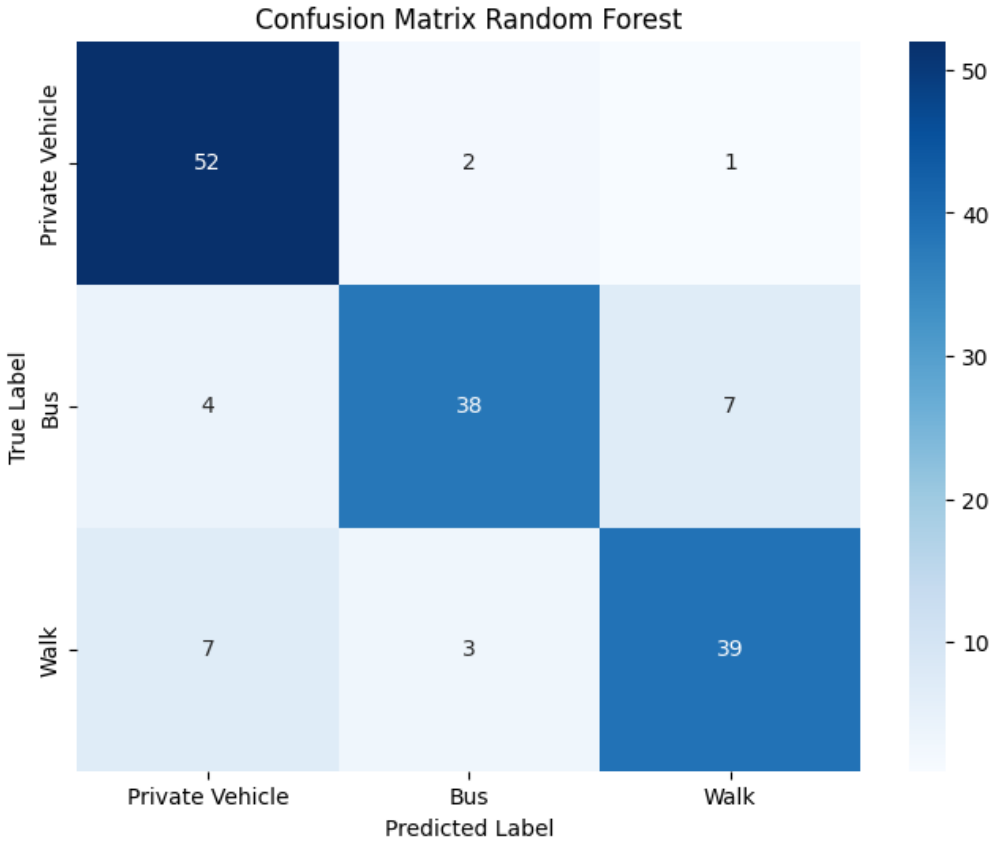


Figure 56. Confusion Matrix - Random forest - Geodesic Distance

The overall accuracy of the model was 0.84, correctly predicting 129 out of 153 instances. Recall for private vehicle was 0.95, retrieving 52 out of 55 relevant instances. Recalls for bus and walk modes were 0.78 and 0.80, retrieving 38 and 39 relevant instances respectively. The highest precision is for bus with 0.88. Swapping distances decreased the overall performance of the model, 0.84 compared to 0.88 overall accuracy of the previous default model. The next step was to tune the parameters of the model attempting to increase the performance. Again, max_depth, number of estimators and max features were configured via grid search. The results are illustrated below.

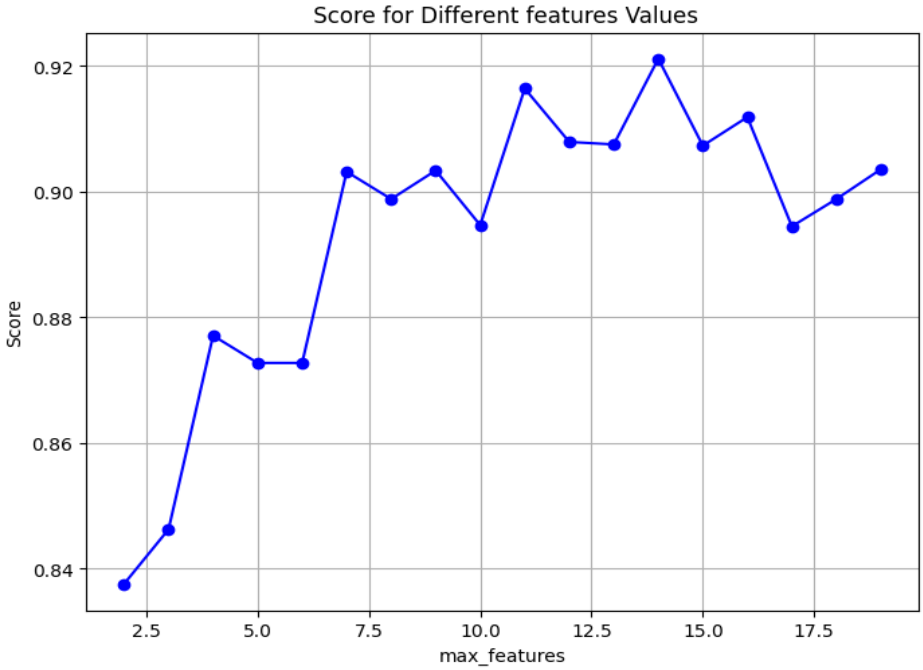


Figure 57. Grid Search RF - Max features - Geodesic Distance

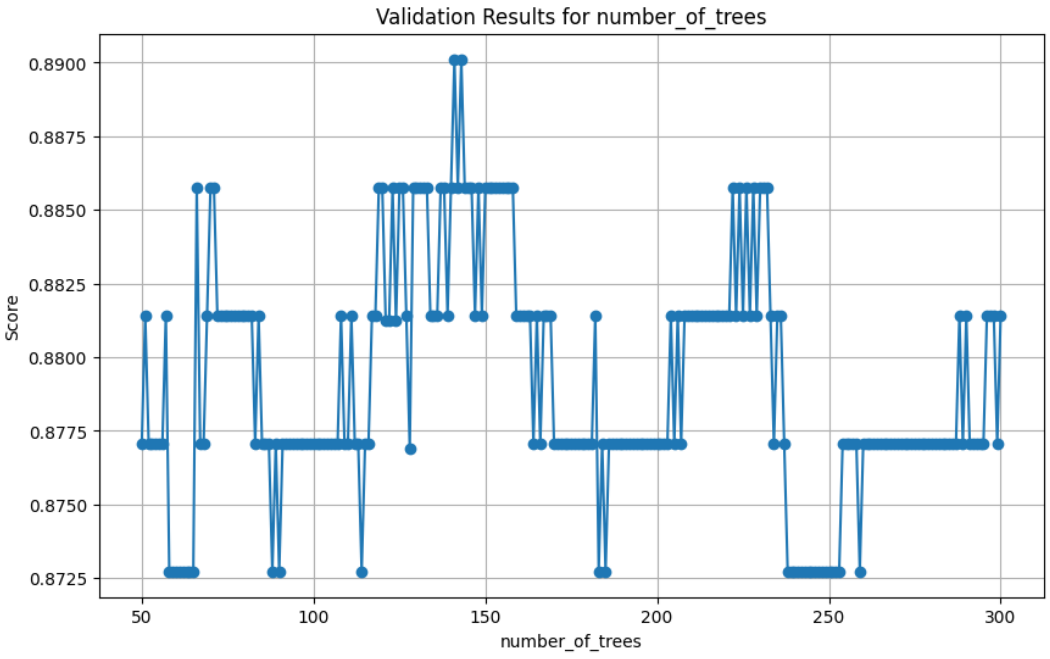


Figure 58. Grid Search RF - Number of Trees - Geodesic Distance

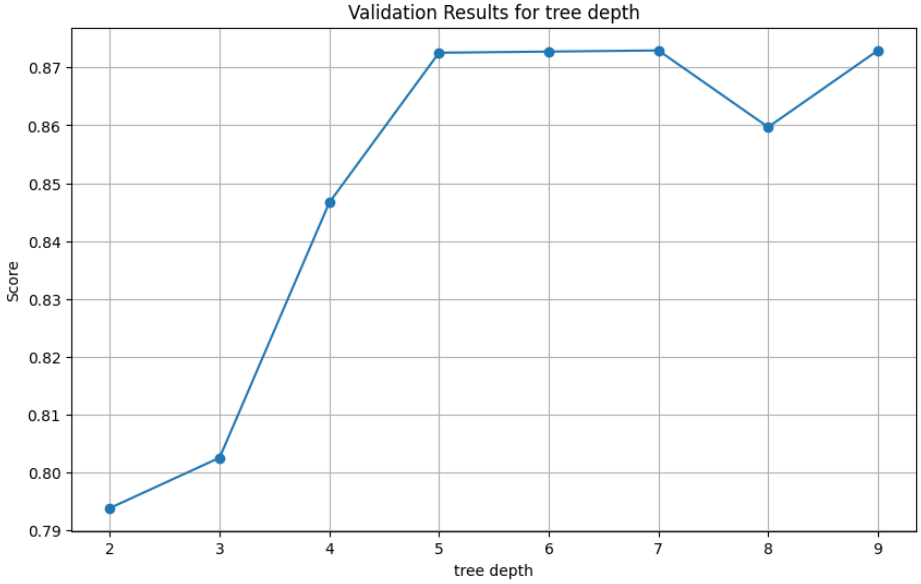


Figure 59. Grid Search RF- Max Depth - Geodesic Distance

Afte finding the optimal areas, a final grid search was configured using the following parameter values.

```
'n_estimators': list(range(125, 142)),
'max_features': [14],
'max_depth': [5, 7]}
```

The best new parameters were found for **max_depth = 5, n_estimators = 141 and max_features = 14**. The out of bag error was at 0.1272 while the **validation accuracy was at 0.90**. The tuned model was then retrieved and evaluated on the test set. The results are depicted below.

Table 12. Tuned Random Forest performance - Geodesic Distance

	Precision	Recall	F1	Accuracy
Private Vehicle	0.88	0.95	0.91	0.86
Bus	0.88	0.78	0.83	
Walk	0.80	0.84	0.82	
Macro Average	0.86	0.85	0.85	

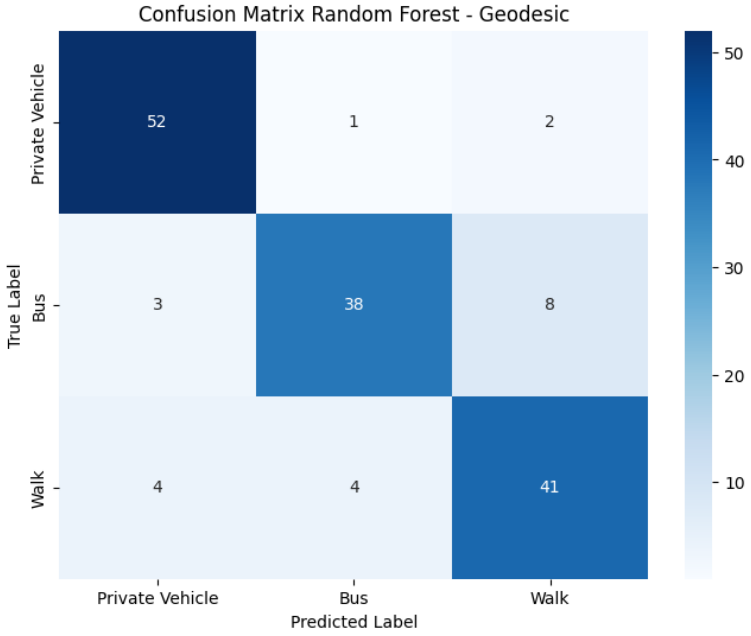


Figure 60. Confusion Matrix - Runed RF - Geodesic Distance

The overall accuracy of the tuned model increased from 0.84 to 0.86. Precision increased for private vehicle from 0.83 to 0.88 while recall remained at 0.95. Recall also increased for walk, from 0.80 to 0.84 while precision dropped from 0.83 to 0.80. All the metrics for bus remained the same. The highest misclassification occurs between walk and bus where the model predicted 8 instances as walk while the true label was bus. The feature importances for the model construction were also extracted. The results are illustrated in the plot below. The trend is likewise the previous Random Forest model, with geodesic distance at the top followed by time and Health, Cost and Income and Weather. The only feature not used in any of the individual trees was whether the respondent was 51 and more or not.

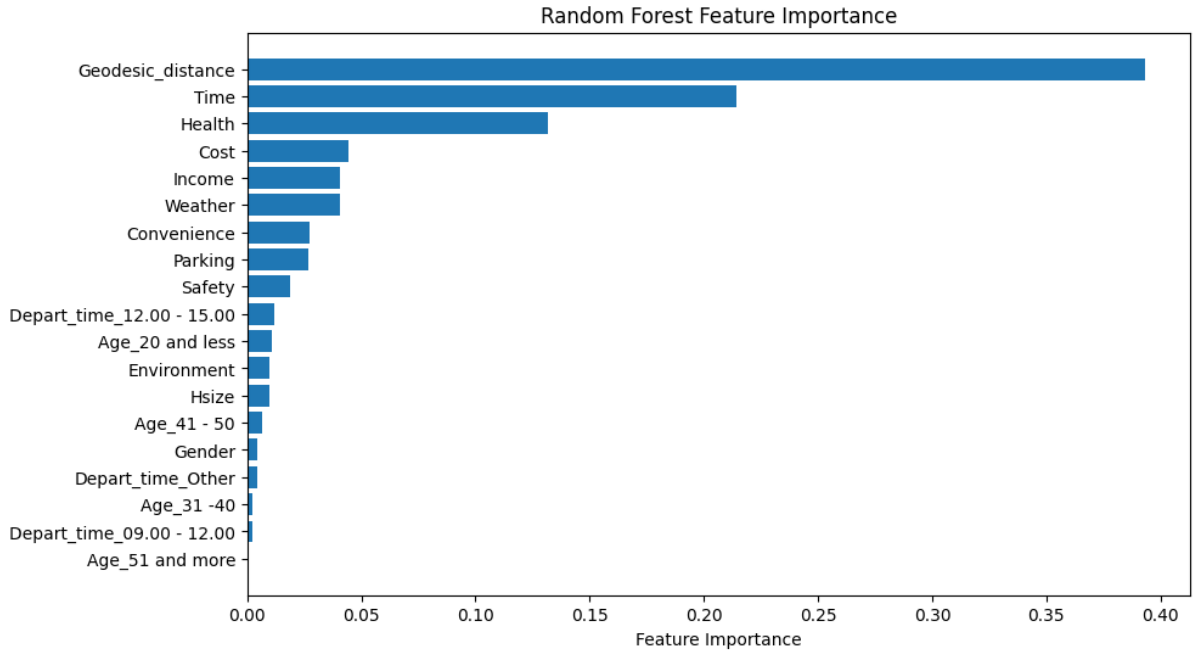


Figure 61. Random Forest feature importance - Geodesic Distance

Table 13. Random Forest (Distance vs Geodesic Distance)

	Random Forest (Distance)			Random Forest (Geodesic distance)		
	Private vehicle	Bus	Walk	Private vehicle	Bus	Walk
Precision	0.96	0.95	0.84	0.88	0.88	0.80
Recall	0.93	0.86	0.96	0.95	0.78	0.84
F1-score	0.94	0.96	0.90	0.91	0.83	0.82
Accuracy	0.92			0.86		
AUC-Macro	0.983			0.96		

The Random Forest model, utilizing the Distance metric, demonstrates better performance in almost every metric. The only exception occurs for private vehicle class, aa the model utilizing geodesic distance demonstrated a recall of 0.95 compared to a recall of 0.93 of the other model. Overall, the first model has a better performance overall, compared to geodesic distance-based model.

4.1.6. XGBoost

The next model applied was that of XGboost. The model was created using the **xgb.XGBClassifier(objective='multi:softmax', num_class=3, random_state=42)** function.

The objective=”multi:softmax” parameter indicates that the task of the model is to classify more than two classes, while the num_class = 3 parameter indicates the number of classes. Similarly to random forest, the random_state parameter is set to achieve reproducibility of the results. The model was then fitted on the training set and consequently evaluated on the test set. The results are depicted below.

Table 14. XGBoost performance - Default

	Precision	Recall	F1	Accuracy
Private Vehicle	0.89	0.98	0.93	0.92
Bus	0.95	0.84	0.89	
Walk	0.92	0.92	0.92	
Macro Average	0.92	0.91	0.91	

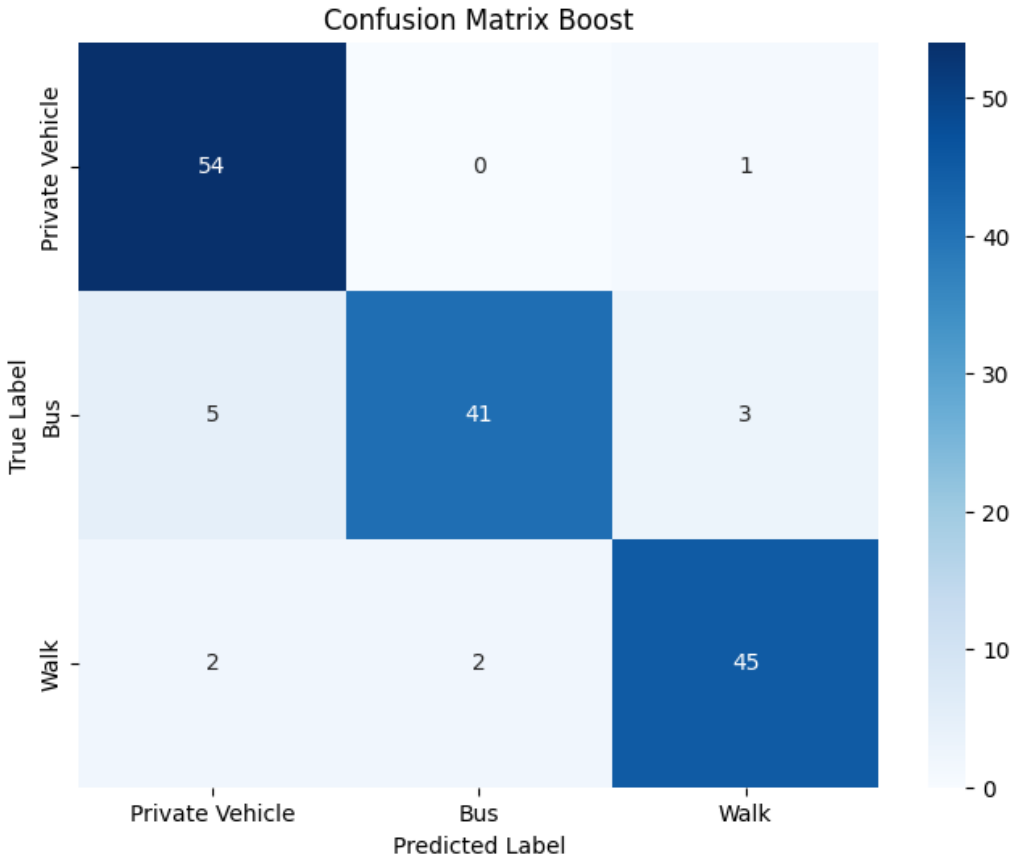


Figure 62. Confusion Matrix - XGBoost - Default

The overall accuracy of the model was 0.92, correctly predicting 140 out of 153 total instances. Recall for private vehicle and walk, was found 0.98 and 0.92 respectively. Bus has the highest precision with 0.95 but a lower recall with 0.84. The highest misclassification occurs between “bus” and “private vehicle”, as the model predicted 5 instances as “private vehicle” while the true label was “bus”. Although the model's performance appears satisfactory, various sets of parameters were modified to assess if performance could be enhanced. The default values of the model are the following: **n_estimators = 100, subsample = 1, colsample_bytree =1 , learning_rate = 0.3**. The number of estimators refers to the number of trees built like Random Forest. The subsample parameter refers to the fraction of the training data that are sampled during the training phase. The colsample_bytree parameter refers to the fraction of features that are sampled for each tree. Finally, the learning_rate parameter refers to the contribution of each tree to the overall model, while affecting how quickly or slowly the model adapts to the patterns in the training data. The max_depth parameter was also tuned for the model. Like random forest and decision tree, tree depth controls the individual tree size. Those parameters were configured through the GridSearchCV() function individually to create a final grid and evaluate the model.

Initially, a grid search was conducted to test the number of estimators within the range of 50 to 300 trees. The outcomes are illustrated below.

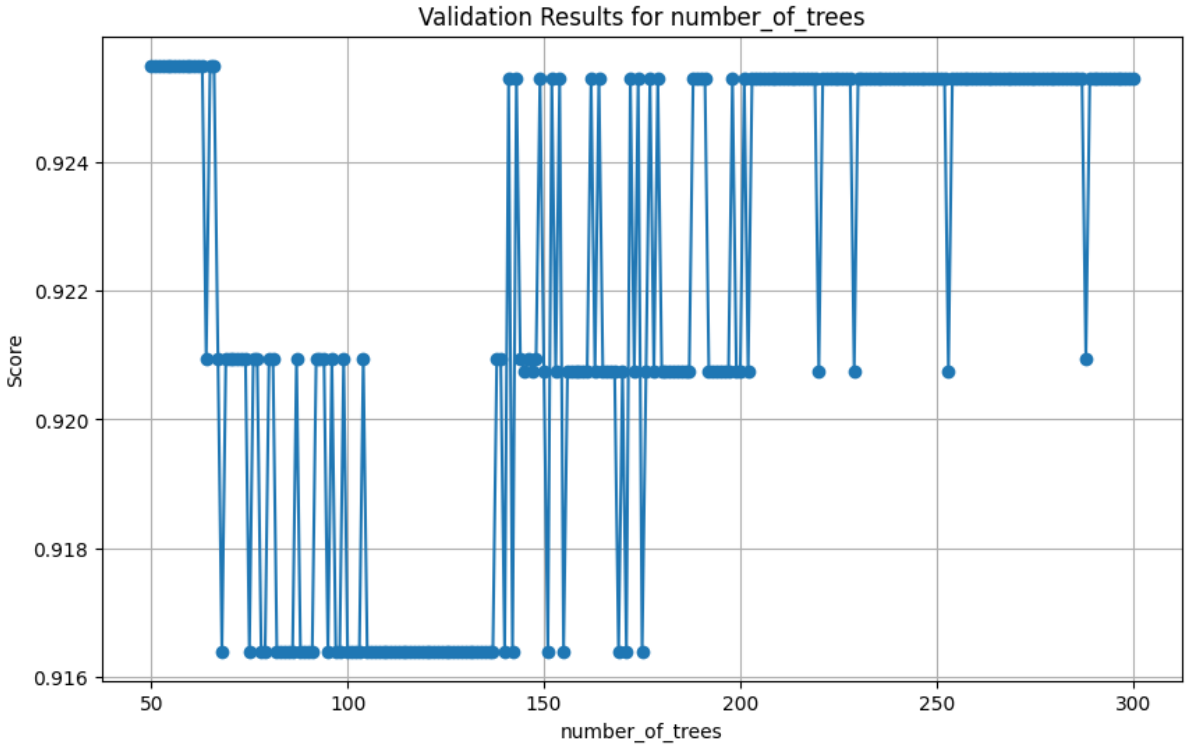


Figure 63. Grid Search XGBoost - Number of Trees

The overall validation accuracy varies between the range, though is higher between 50 and 70 number of trees. The next parameter that was tested was that of learning_rate with values from 0.1 up to 1. The results are illustrated below.

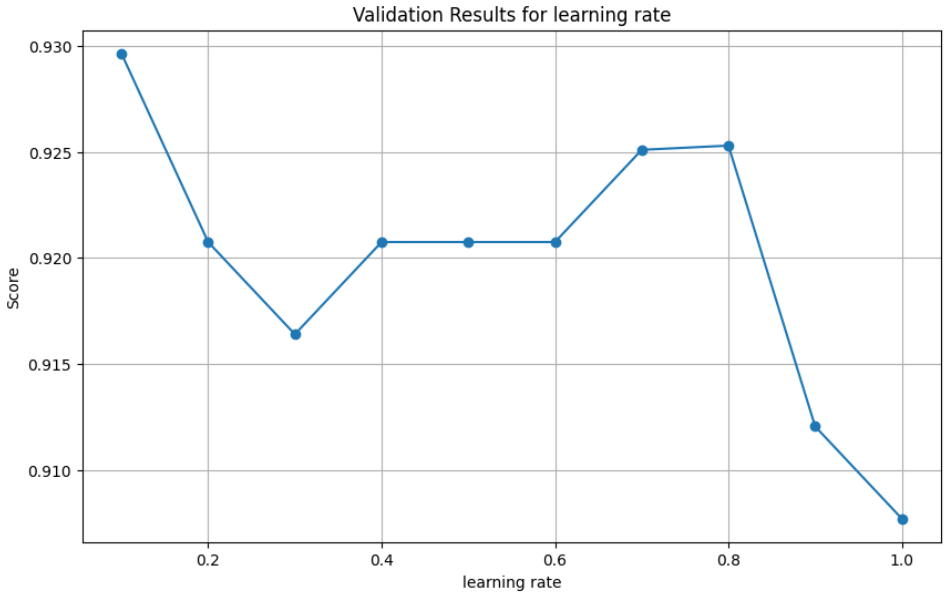


Figure 64. Grid Search XGBoost - Learning rate

The results indicate best validation accuracy for learning_rate = 0.1. The next parameter tested was subsamples, with values from 0.1 to 1. The results are illustrated below.

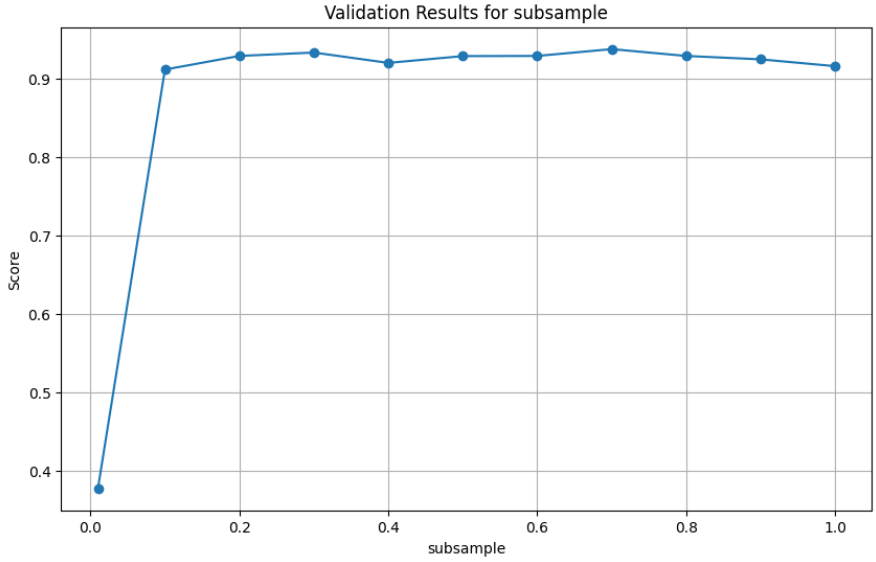


Figure 65. Grid Search XGBoost - Subsample

It is identified that the validation accuracy is best for 0.7. Though, validation accuracy is consistently high for all values of subsample. The next parameter tuned was that of colsample_bytree. The results are illustrated below.

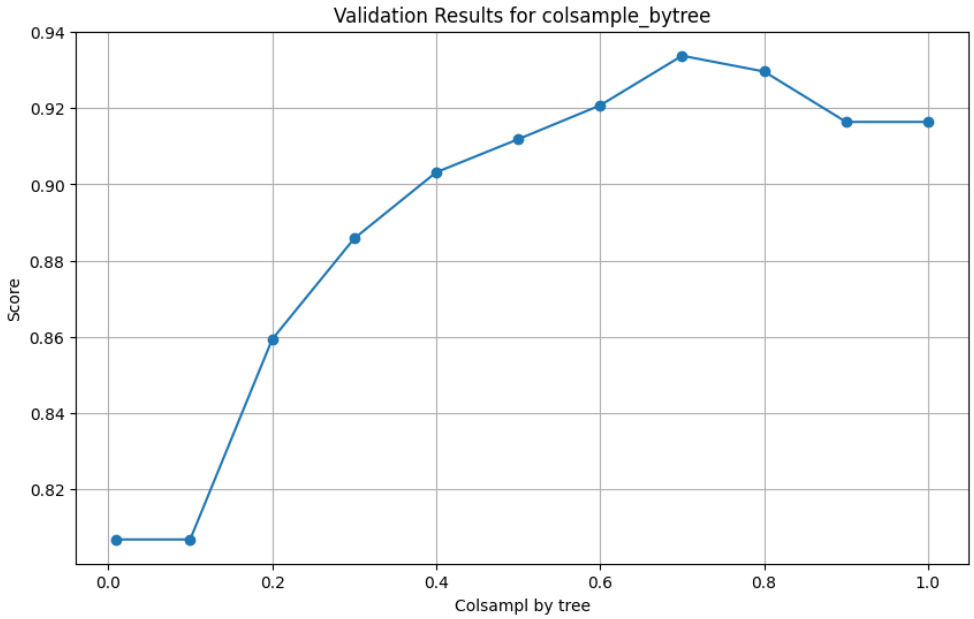


Figure 66. Grid Search XGBoost - Colsample

The validation accuracy is best for colsample value of 0.7. The last parameter tuned was that of max_depth. The results are depicted below.

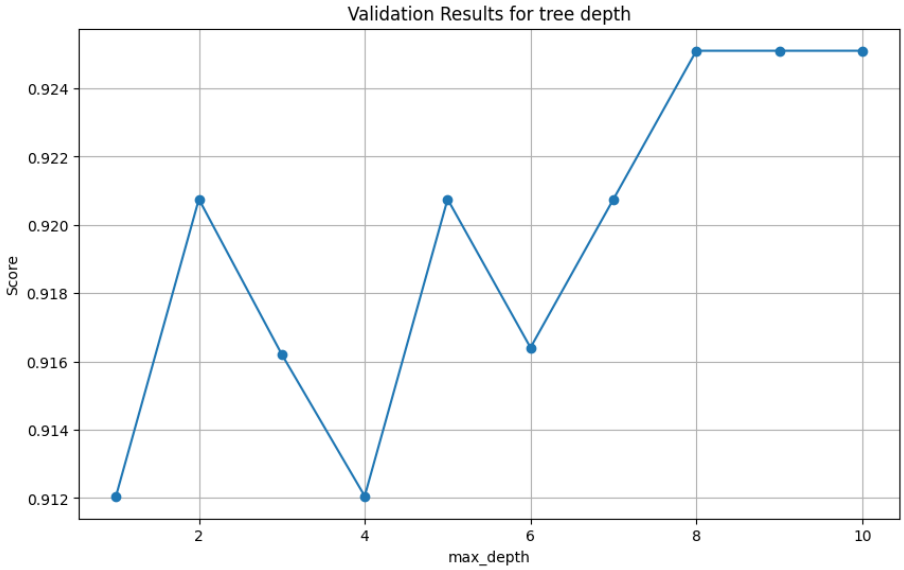


Figure 67. Grid Search XGBoost - max depth

Accuracy is higher for a max depth of 8, 9 and 10, though the parameter will be set at 8 to limit the growth of the trees.

After finding the optimal values a final grid was run to tune the model. The parameter values that were tuned through grid search are illustrated below.

```
'n_estimators': list(range(50, 101)),
'subsample': [0.7],
'colsample_bytree' : [0.7, 0.8],
'learning_rate': [0.1],
'max_depth':[8]}
```

The optimal parameters values from grid search are the following: **n_estimators = 50, subsample = 0.7, colsample_bytree = 0.8, learning_rate = 0.1, max_depth=8 and validation accuracy of 0.946**. The tuned model was then retrieved and evaluated on the test set. The results are depicted below.

Table 15. Tuned XGBoost performance

	Precision	Recall	F1	Accuracy
Private Vehicle	0.95	0.96	0.95	0.93
Bus	0.96	0.88	0.91	
Walk	0.88	0.94	0.91	
Macro Average	0.93	0.93	0.93	

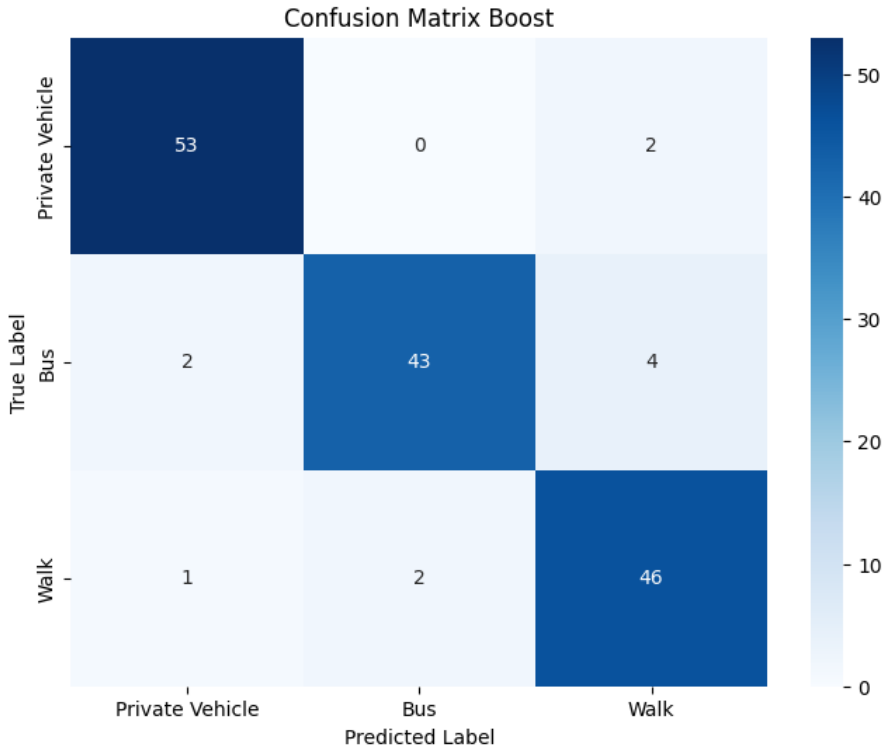


Figure 68. Confusion Matrix Tuned XGBoost

The overall accuracy of the tuned model increased from 0.92 to 0.93. Precision for private vehicle increased from 0.89 to 0.95 with a trade off in recall as it decreased from 0.98 to 0.95. Recall for bus increased from 0.84 to 0,88, while precision slightly increased from 0.95 to 0.96. Recall also increased for walk, from 0.92 to 0.94 while precision decreased from 0.92 to 0.88. The highest misclassification occurred between walk and bus, as the model predicted 4 cases as walk while the true label was bus. Overall, the tuned model demonstrates a more balanced performance compared to the default one. The feature importances were also extracted for the model. The results are depicted below.

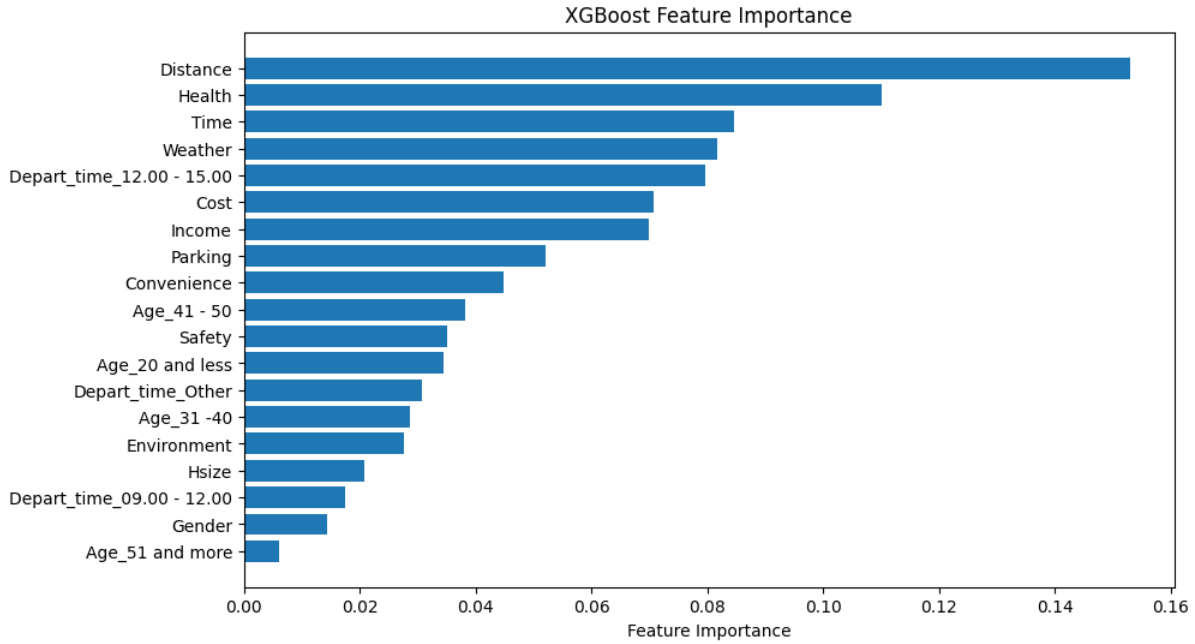


Figure 69. XGBoost feature importance

Like the previous tree-based models, distance remains the most important feature for constructing the individual trees. Surprisingly, while time was the second most important feature for all the previous models, XGboost ranks influence of physical activity and health as the second most important feature. High on the list is also one of the binary features and whether the respondent commutes to work between 12.00 – 15.00 or not. At the bottom of the list are Gender and whether the age of the commuter was more than 51 years old or not. Still, the model used all the features for the individual tree construction. The last experimentation was again swapping Distance with Geodesic_Distance. A new model was created, trained on the train set and consequently evaluated on the test set. The results are illustrated below.

Table 16. XGBoost - Geodesic Distance - Default

	Precision	Recall	F1	Accuracy
Private Vehicle	0.85	0.96	0.91	0.86
Bus	0.87	0.82	0.84	
Walk	0.87	0.80	0.83	
Macro Average	0.86	0.86	0.86	

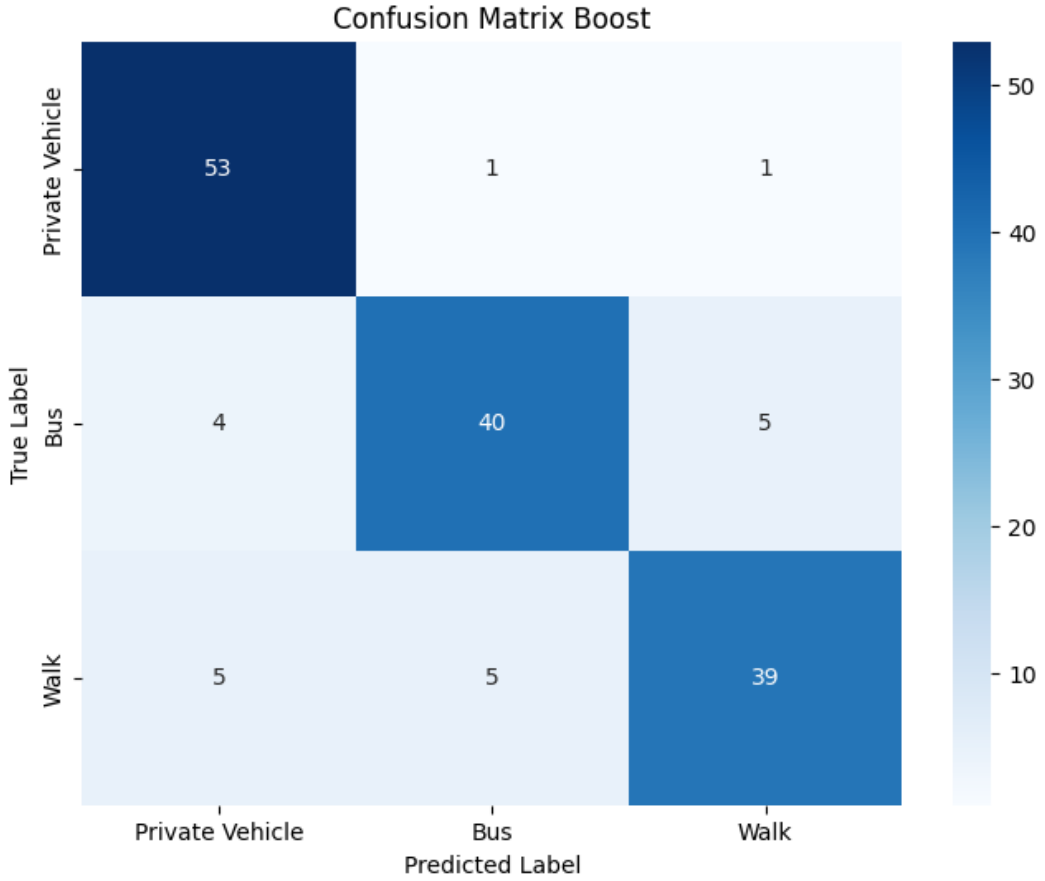


Figure 70. Confusion Matrix - XGBoost -Geodesic Distance

The overall accuracy of the model dropped from 0.92 to 0.86 by swapping distances. Recall for private vehicle was 0.96, followed by bus at 0.82 and walk with 0.80. Precision is higher for both bus and walk with 0.87. While the overall metrics of the model dropped swapping distances, the parameters were re-tuned to increase the performance of the model. The same parameters as before were tuned to find optimal areas. The validation results are illustrated below.

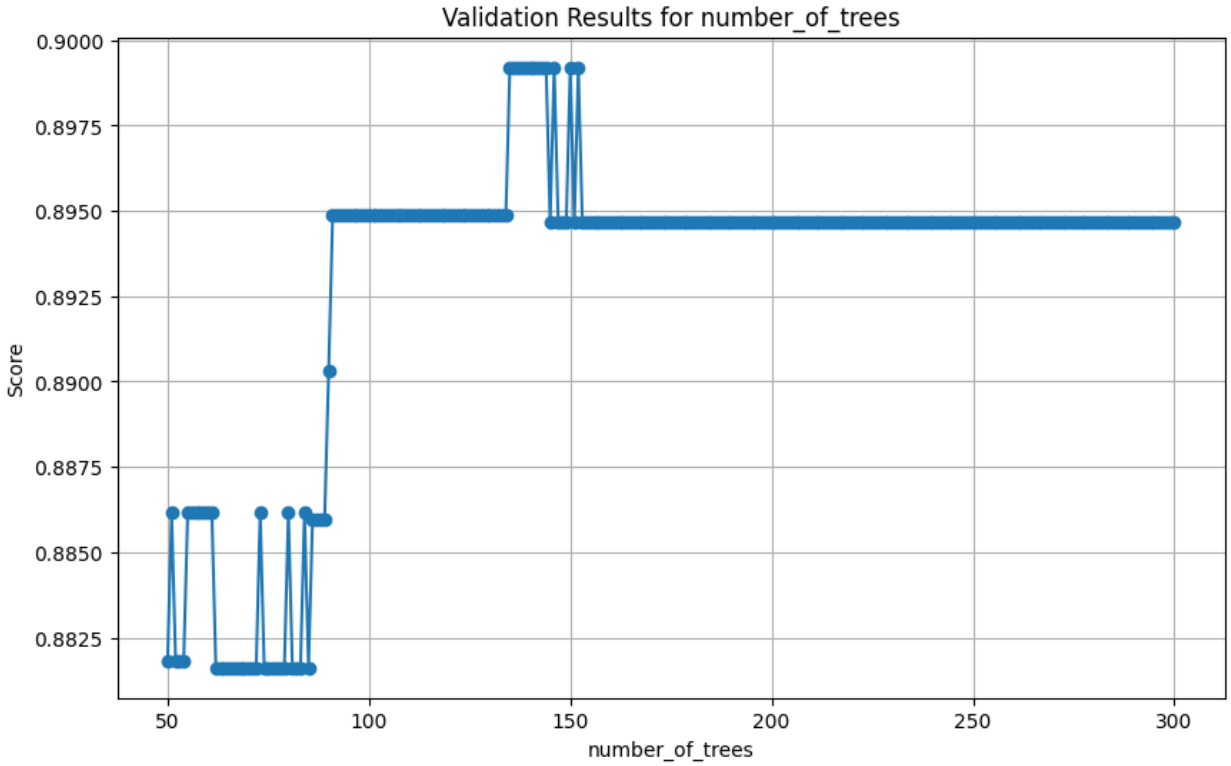


Figure 71. Grid Search XGB - Number of Trees - Geodesic Distance

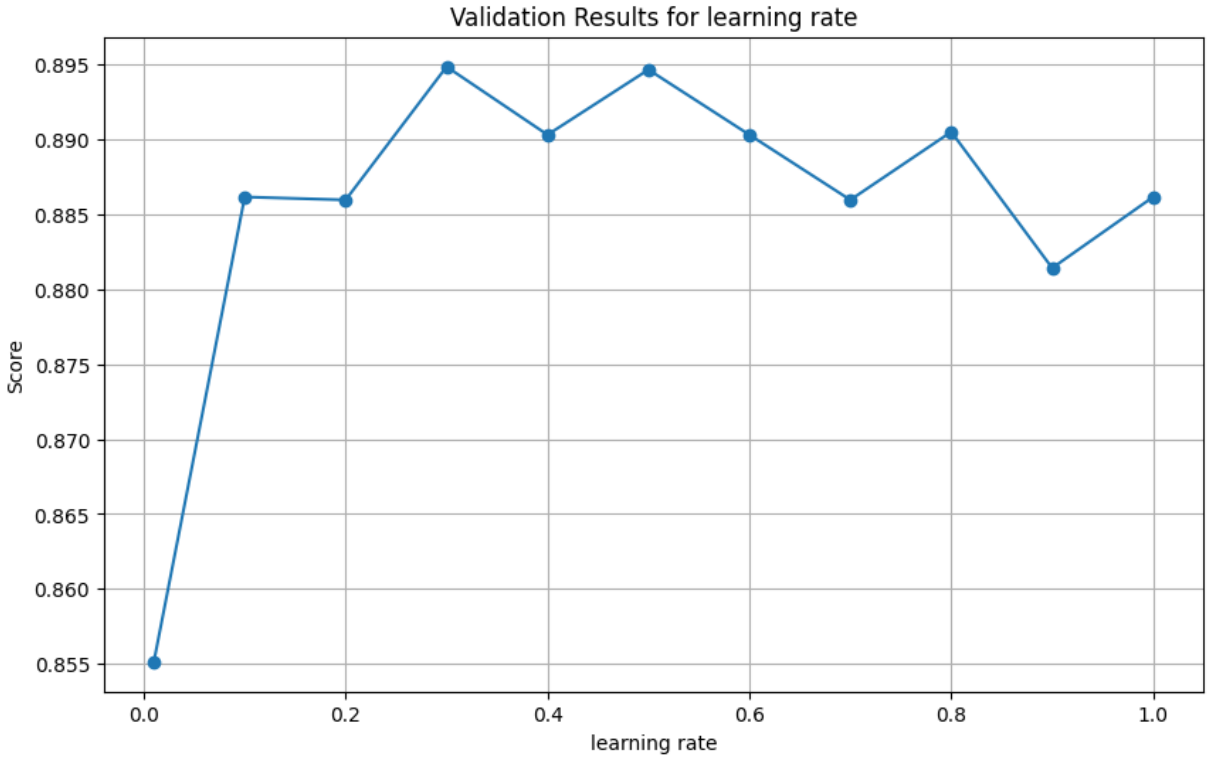


Figure 72. Grid Search XGB - Learning rate - Geodesic Distance

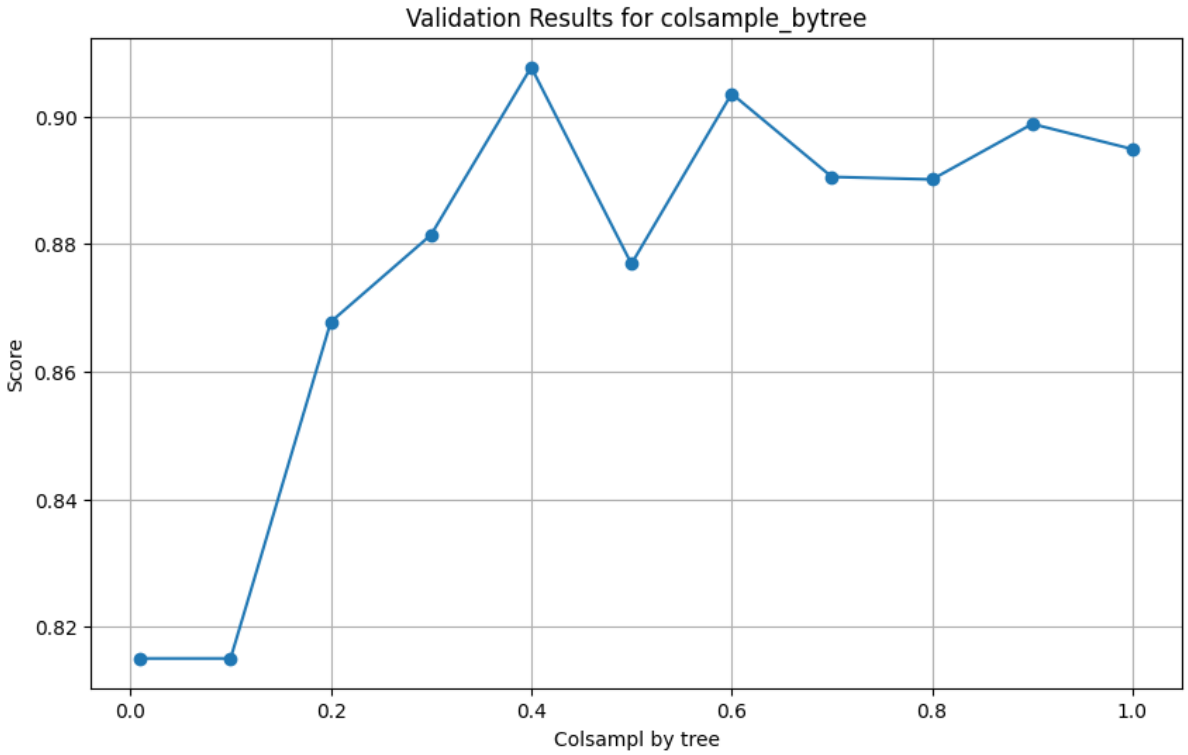


Figure 73. Grid Search XGB - Colsample - Geodesic Distance

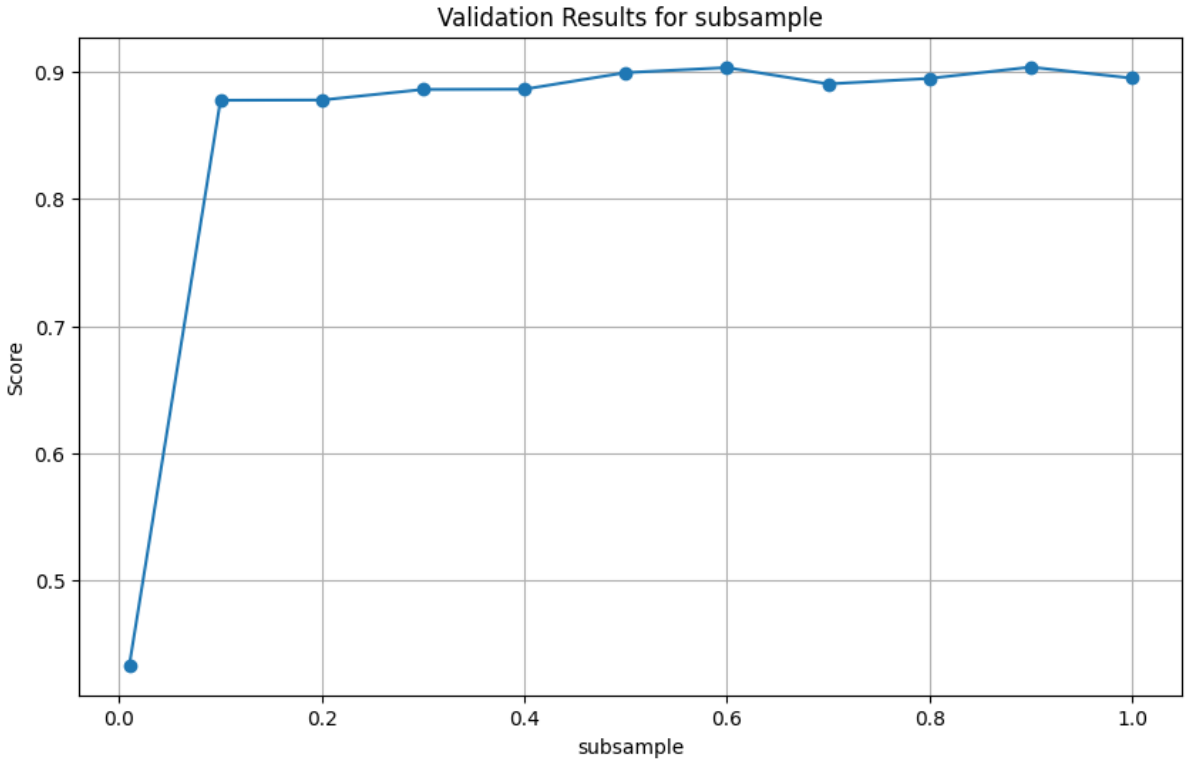


Figure 74. Grid Search XGB - Subsample - Geodesic Distance

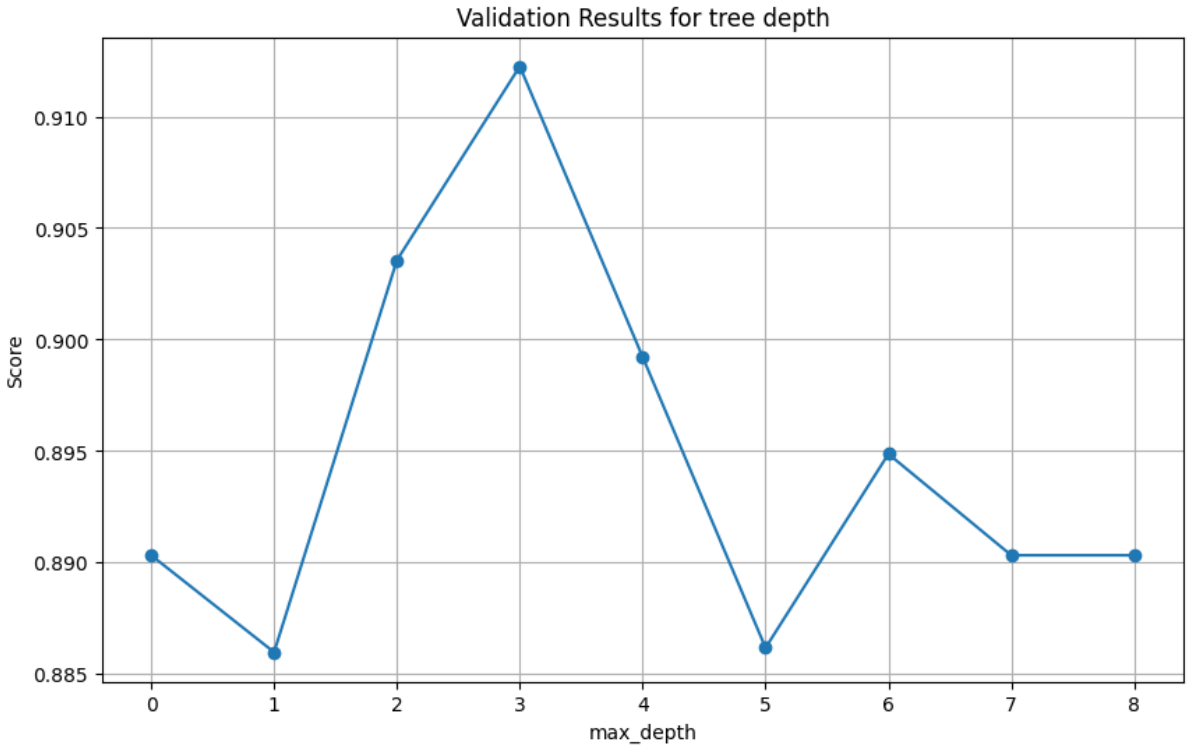


Figure 75. Grid Search XGB - Max depth - Geodesic Distance

After finding optimal areas, a final grid search was run with the following parameters.

```
param_grid = {
    'n_estimators': list(range(125, 150)),
    'subsample': [0.6],
    'colsample_bytree' : [0.4],
    'learning_rate': [0.3],
    'max_depth':[3]}
```

The best parameters were found for **colsample = 0.4, learning_rate=0.3, max_depth = 3, n_estimators=126, subsample = 0.6 and a validation accuracy of 0.8903**. The model was then retrieved and evaluated on the test set. The results are illustrated below.

Table 17. Tuned XGBoost performance - Geodesic Distance

	Precision	Recall	F1	Accuracy
Private Vehicle	0.87	0.98	0.92	0.88
Bus	0.91	0.84	0.87	
Walk	0.85	0.80	0.82	
Macro Average	0.88	0.87	0.87	

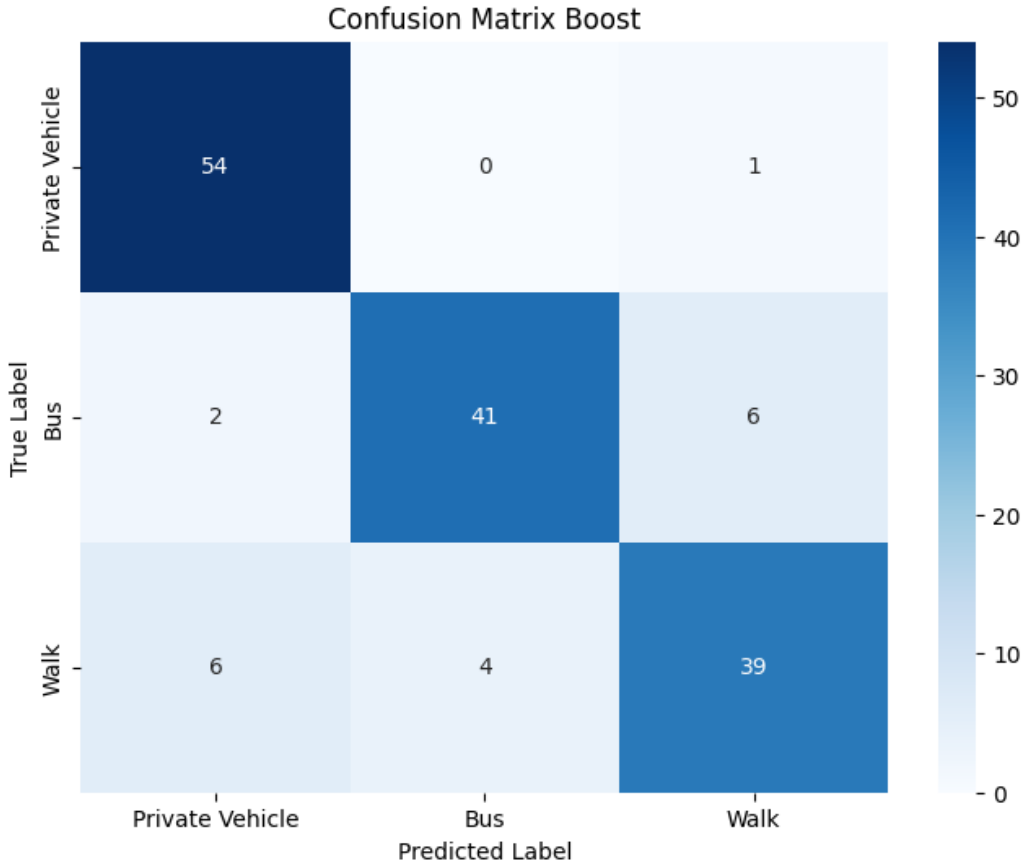


Figure 76. Confusion Matrix Tuned XGBoost - Geodesic Distance

For the tuned model, recall for both bus and private vehicle increased at 0.84 and 0.98 respectively. Precision has increased for both private vehicle and bus from 0.85 and 0.87 up to 0.87 and 0.91 respectively. In contrast, precision for walk decreased from 0.87 to 0.85, while recall remained the same. The model misclassified 6 cases each as walk and private vehicle while their true label was bus and walk respectively. The feature importance was also retrieved for the tuned model. The results are depicted below.

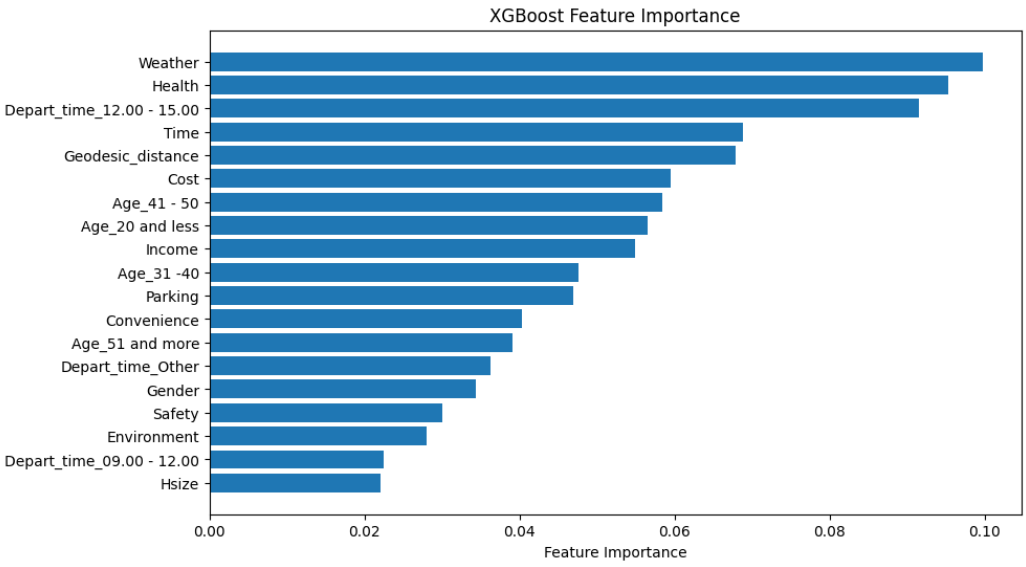


Figure 77. XGBoost feature importance - Geodesic Distance

Surprisingly, the model diverges significantly from the previous models as geodesic distance has dropped at 5th among the most important features. First are now Weather and Health features followed by whether the commute was between 12.00-15.00 or not. The least important features are household size and whether the commute time was between 09.00 – 12.00 or not. While the model seemed to use all the features for the construction of the individual trees, not a concrete decision can be made on the importance below the third place and Depart time (12.00-15.00), since the values for feature importances are relatively close to each other and more data are required. The table below depicts the metrics for both models (Distance vs Geodesic).

Table 18 XGBoost (Distance vs Geodesic Distance)

	XGBoost (Distance)			XGBoost (Geodesic distance)		
	Private vehicle	Bus	Walk	Private vehicle	Bus	Walk
Precision	0.95	0.96	0.88	0.87	0.91	0.85
Recall	0.96	0.88	0.94	0.98	0.84	0.80
F1-score	0.95	0.91	0.91	0.92	0.87	0.82
Accuracy	0.93			0.88		
AUC-Macro	0.98			0.96		

Utilizing Distance instead of Geodesic distance, demonstrates a better overall performance for XGBoost in almost every metric for the individual classes. The only exception is to recall for private vehicle class, decreased from 0.98 to 0.96, though precision is significantly higher, 0.95 compared to 0.87. The AUC macro score is also higher for the first model demonstrating a better predictive performance.

4.1.7. Stacked Model

The last model applied was a Stacked Model. Stacking occurs when multiple models are combined to create an enhanced classifier. For the stacked model to be constructed, base classifiers and a "meta" learner need to be defined. The tree-based models (Decision Tree, Random Forest and XGBoost) were used as the base estimators while Logistic Regression (one vs rest approach) was used as the final estimator. The StackingCVClassifier of mlxtend library was used to create the model. As a first experimentation the default values for all estimators were used. The configuration of the model is illustrated below.

```
base_classifier1 = RandomForestClassifier(random_state=1,n_jobs=-1)
base_classifier2 = xgb.XGBClassifier(objective='multi:softmax',
num_class=3, random_state=42)
base_classifier3 = DecisionTreeClassifier(random_state=2)
meta_classifier = LogisticRegression(multi_class='ovr')
stacking_classifier = StackingCVClassifier(
    classifiers=[base_classifier1, base_classifier2, base_classifier3],
    meta_classifier=meta_classifier,
    cv=10,
    stratify=True,
    random_state=10
)
```

The model was initially fitted on the training set and was evaluated on the test set afterwards. The results from the test set are illustrated on the classification report and the confusion matrix below.

Table 19. Stacked model performance

	Precision	Recall	F1	Accuracy
Private Vehicle	0.83	0.96	0.89	0.88
Bus	0.92	0.73	0.82	
Walk	0.90	0.92	0.91	
Macro Average	0.88	0.87	0.87	

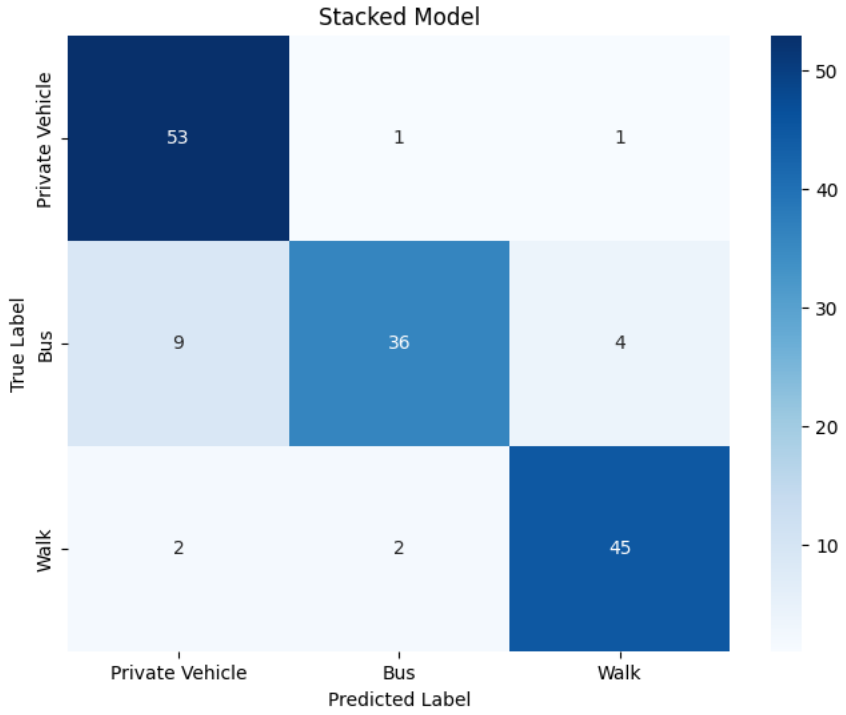


Figure 78. Confusion Matrix - Stacked model

The overall accuracy of the model was 0.88. For "Private Vehicle," the model achieves high recall (0.96) and a precision of 0.83. For bus label, precision is at 0.92, while recall is low at 0.73 retrieving 36 out of 49 true labels. For walk, precision and recall are 0.90 and 0.92 respectively, retrieving 45 out of 49 relevant instances. The highest misclassification occurs between bus and private vehicle, where the model predicted 9 instances as private vehicle while the true label was bus.

The second experimentation was to use the tuned models from the previous chapters. The configuration of the new stacked model is illustrated below.

```
base_classifier1 = RandomForestClassifier(random_state=1,n_jobs=-1, max_depth=7, max_features=10, n_estimators=75)
base_classifier2 = xgb.XGBClassifier(objective='multi:softmax', num_class=3, random_state=1,
    n_estimators=50,
    subsample=0.7,
    colsample_bytree=0.8,
    learning_rate=0.1,
    max_depth=8)
base_classifier3 = DecisionTreeClassifier(random_state=21, max_depth=5, min_samples_leaf = 3, min_samples_split= 2)
meta_classifier = LogisticRegression(multi_class='ovr')
stacking_classifier = StackingCVClassifier(
    classifiers=[base_classifier1, base_classifier2, base_classifier3],
    meta_classifier=meta_classifier,
    cv=10,
    stratify=True,
    random_state=10
)
```

The model was initially fitted on the training set and was evaluated on the test set afterwards. The results from the test set are illustrated on the classification report and the confusion matrix below.

Table 20. Stacked Model performance (optimal estimators)

	Precision	Recall	F1	Accuracy
Private Vehicle	0.96	0.93	0.94	0.90
Bus	0.86	0.86	0.86	
Walk	0.86	0.90	0.88	
Macro Average	0.89	0.89	0.89	

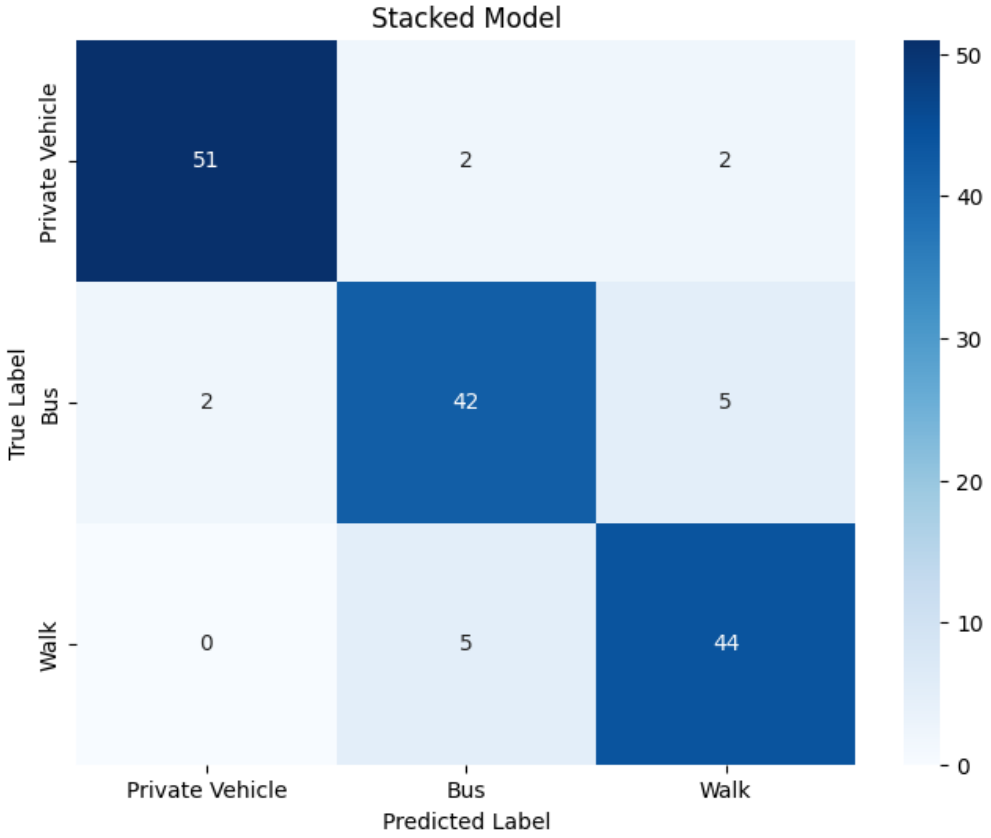


Figure 79. Confusion Matrix Stacked (optimal estimators)

The overall accuracy of the new stacked increased from 0.88 to 0.90. There is a trade off in recalls for the classes. For bus it significantly increased from 0.73 to 0.86, while for private vehicle and walk it slightly dropped at 0.93 and 0.90 from 0.96 and 0.92 respectively. Precision for private vehicle also increased from 0.83 to 0.96, while for bus and walk it dropped at 0.86. The biggest misclassification occurs between walk and bus, where the model predicted 5 instances as walk when the true label was bus and vice versa. The coefficients of the meta estimator (Logistic Regression) were retrieved for the model. The output is illustrated below.

Table 21. Meta estimator coefficients

	Random Forest	XGBoost	Decision Tree
Private vehicle	-2.50	-0.96	-0.85
Bus	0.65	-0.31	-0.35
Walk	1.73	1.72	1.39

Generally, the coefficients represent the change in log odds for predicting that class for a one-unit increase. Though, while in a standard Logistic regression model those coefficients are assigned to the features of the model, in the stacked version those values are assigned to each base estimator. Higher values indicate a higher change in the log odds of the class. For instance, an increase in the Random Forest prediction for private vehicle lead to a decrease of the log odds for that class by 2.5. Additionally,

an increase in XGboost prediction for class walk, increase the log odds for that class by 1.72. The same procedure could be implemented for every coefficient.

As a second experimentation, again geodesic distance was swapped with Distance and a new stacked model was constructed using the optimal parameters of models utilizing geodesic distance. The configuration for the model is illustrated below.

```
base_classifier1 = RandomForestClassifier(random_state=1,n_jobs=-1, max_depth=5, max_features=14, n_estimators=141)
base_classifier2 = xgb.XGBClassifier(objective='multi:softmax', num_class=3, random_state=42,
                                     colsample_bytree= 0.4, learning_rate= 0.3, max_depth= 3, n_estimators= 126, subsample= 0.6)
base_classifier3 = DecisionTreeClassifier(random_state=21, max_depth=6, min_samples_leaf = 2, min_samples_split= 2)
meta_classifier = LogisticRegression(multi_class='ovr')
stacking_classifier = StackingCVClassifier(
    classifiers=[base_classifier1, base_classifier2, base_classifier3],
    meta_classifier=meta_classifier,
    cv=10,
    stratify=True,
    random_state=1
)
```

Table 22. Stacked - Geodesic Distance

	Precision	Recall	F1	Accuracy
Private Vehicle	0.87	0.96	0.91	0.86
Bus	0.87	0.80	0.83	
Walk	0.85	0.82	0.83	
Macro Average	0.86	0.86	0.86	

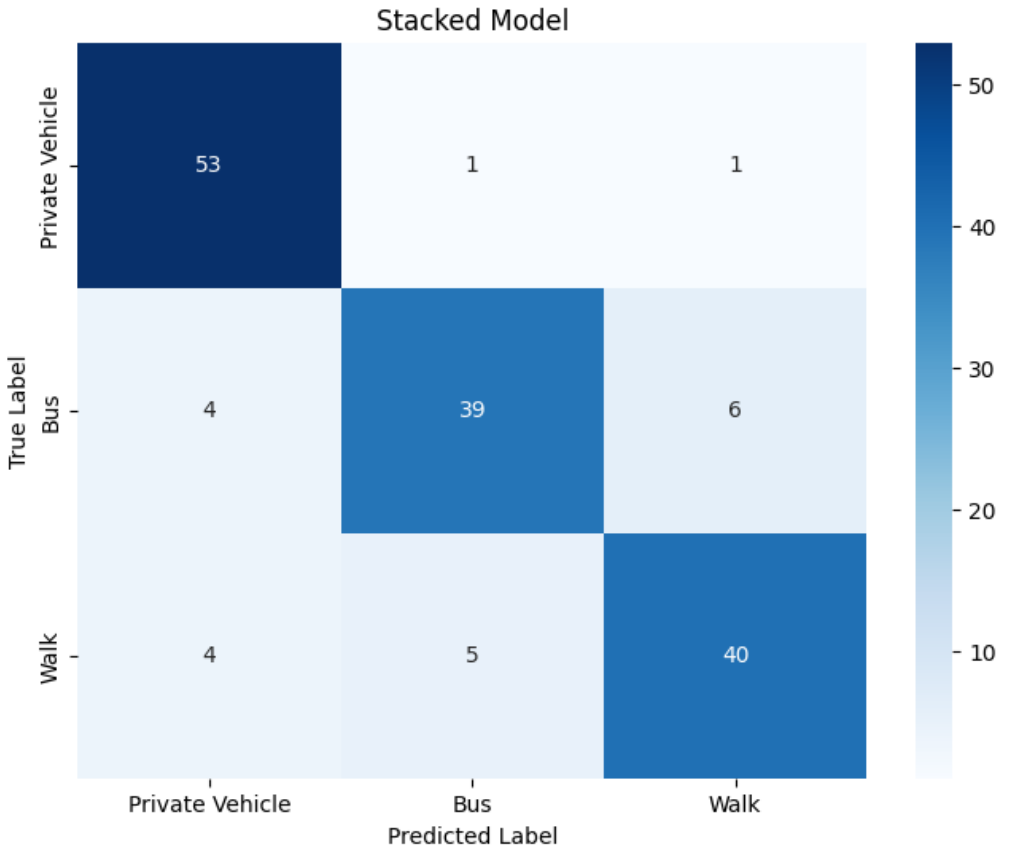


Figure 80. Confusion Matrix Stacked - Geodesic Distance

The overall accuracy of the model was 0.86. For "Private Vehicle," the model achieves high recall (0.96) and a precision of 0.87. For "Bus" category, precision is at 0.87, while recall is lower at 0.80. For walk, precision and recall are also lower at 0.85 and 0.82 respectively. The highest misclassification occurs between walk and bus classes as the model predicted 6 cases as walk while the true label was bus. The overall performance of the model dropped by swapping distances.

The table below depicted the metrics of the models utilizing different distance metrics.

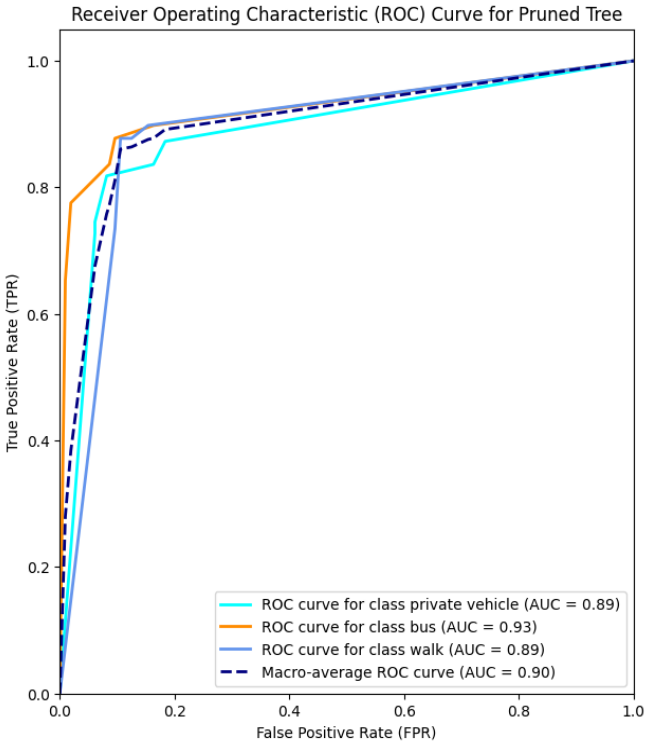
Table 23. Stacked Model (Distance vs Geodesic Distance)

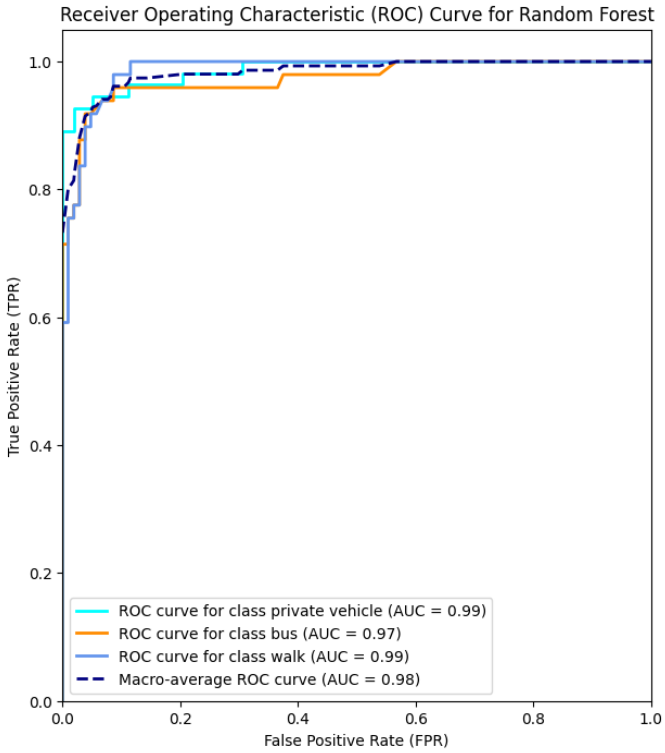
	Stacked (Distance)			Stacked (Geodesic distance)		
	Private vehicle	Bus	Walk	Private vehicle	Bus	Walk
Precision	0.96	0.86	0.86	0.87	0.87	0.85
Recall	0.93	0.86	0.90	0.96	0.80	0.82
F1-score	0.94	0.86	0.88	0.91	0.83	0.83
Accuracy	0.90			0.86		
AUC-Macro	0.9304			0.9265		

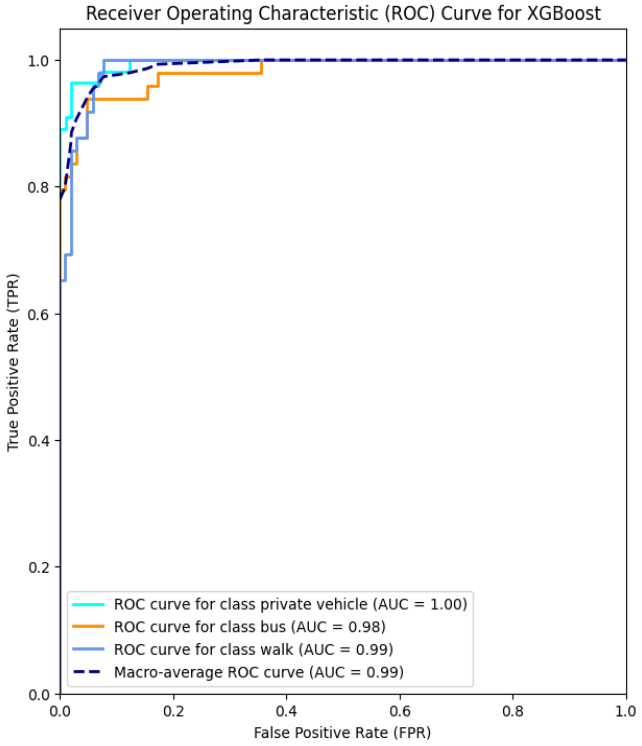
The first model demonstrates a more balanced performance across the classes, while the auc macro score is slightly higher that the model with Geodesic distance, indicating a better predictive capability.

4.1.8. Model Comparison

Before assessing the model performance, the ROC curves for the best models were constructed. The ROC curve is a graphical representation that evaluates the trade-off between true positive rate and false positive rate. The curve is generated by calculating true positive rate and false positive rate values across various probability thresholds. Although these curves are typically created for binary classification scenarios, in a multi-classification task, curves are generated for each individual class using a "One vs Rest" approach. For example, in the case of the "private vehicle" class, the true positive rate signifies the ratio of correctly predicted observations belonging to that class. Conversely, the false positive rate for "private vehicle" indicates the ratio of observations falsely predicted as belonging to that class when they actually belong to any of the other classes (bus or walk). The metric representing the model's performance is the area under the curve (AUC), calculated for each class individually in a multiclass task. The AUC macro score is then derived by averaging the AUC scores for each class. This score ranges from 0 to 1, providing an overall assessment of the model's performance. The ROC curves of the models utilizing Distance feature are presented below.







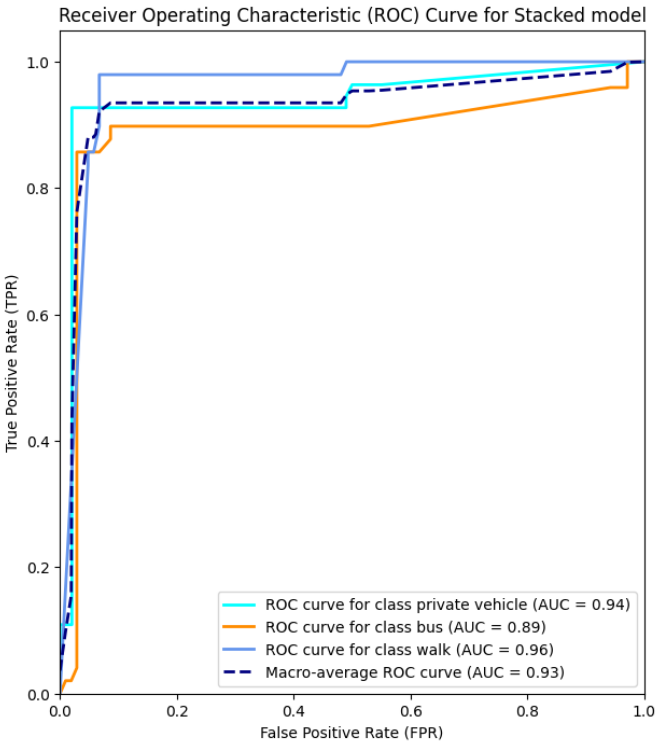


Figure 81. ROC Curves - Distance

After applying the 4 models to the training set and evaluating their predicting performance on the test set, while also constructing their ROC curves, a concrete decision must be made on which model performs the best and can be trusted for prediction on new unseen data. For that purpose, a comparative table was constructed highlighting the averages of Precision, Recall, F-score, AUC between classes and the overall accuracy of the best models that were constructed. The models below all use the Distance feature.

Table 24. Model Comparison - Distance

Models with Distance	**Decision Tree**	**Random Forest**	**XGBoost**	**Stacked Model**
Precision (Average)	0.83	0.92	**0.93**	0.89
Recall (Average)	0.83	0.91	**0.93**	0.89
F-Score (Average)	0.83	0.91	**0.93**	0.89
AUC (Average)	0.90	0.98	**0.985**	0.93
Accuracy	0.83	0.92	**0.93**	0.90

From the table above it is evident that XGBoost and Random Forest demonstrate the best overall performance for Precision, Recall and F scores. They also demonstrate the highest macro AUC above 0.98. XGBoost demonstrates a slightly better performance as AUC for XGBoost is 0.986 compared to 0.980 for Random Forest. Hence, XGBoost is the model to be used for new unseen data using the Distance feature.

The roc curves were also generated for the models that utilize Geodesic Distance. The plots are illustrated below along with the model comparative table.

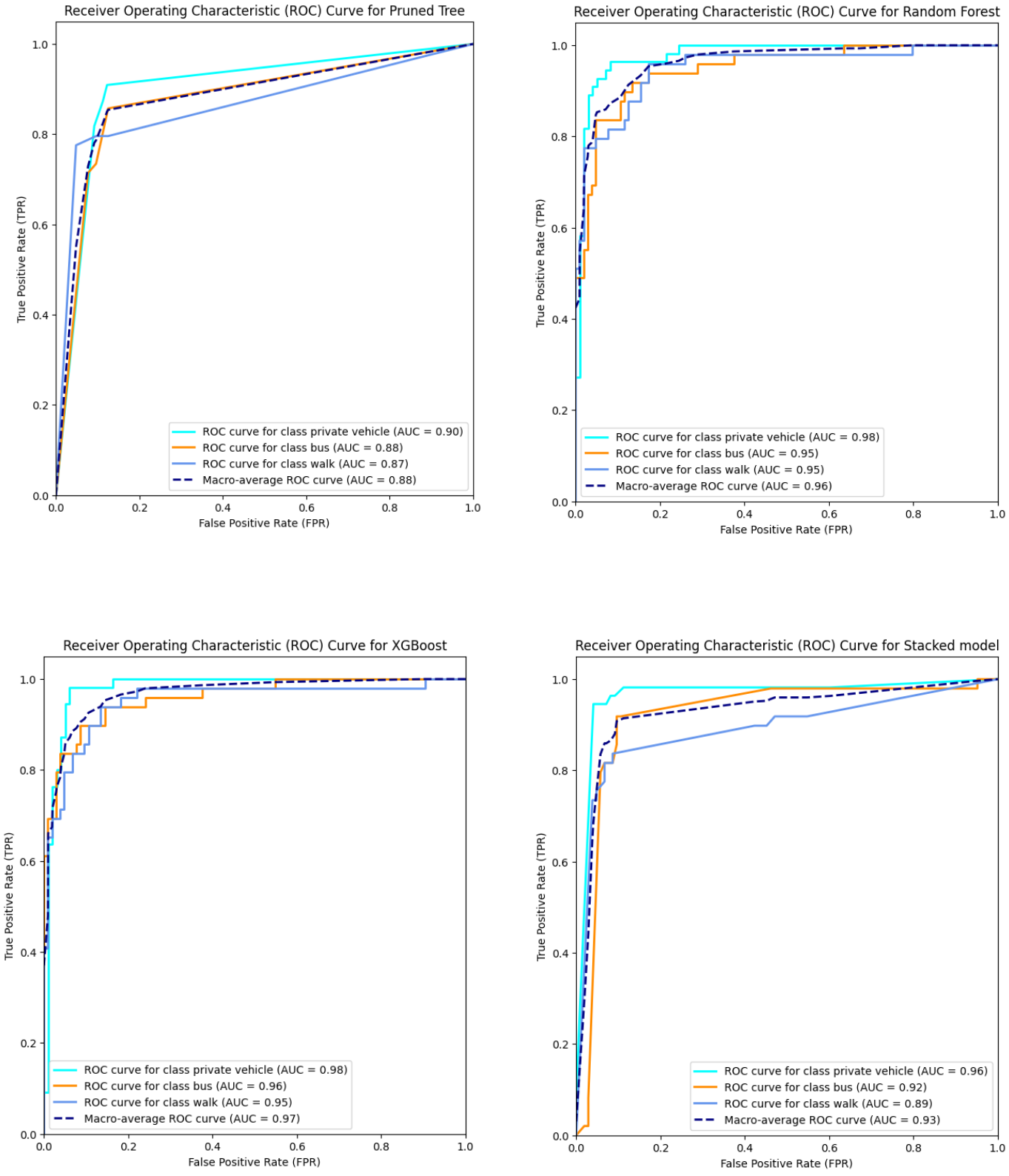


Figure 82. ROC Curves - Geodesic Distance

Table 25. Model Comparison - Geodesic Distance

Models with Geodesic Distance	**Decision Tree**	**Random Forest**	**XGBoost**	**Stacked Model**
Precision (Average)	0.83	0.86	**0.88**	0.86
Recall (Average)	0.82	0.85	**0.87**	0.86
F-Score (Average)	0.82	0.85	**0.87**	0.86
AUC (Average)	0.88	0.96	**0.96**	0.926
Accuracy	0.82	0.86	**0.88**	0.86

From the table above it is evident that again XGBoost demonstrates the best overall performance for Precision, Recall and F scores. It also demonstrates the highest macro-AUC at 0.96. While all the metrics dropped compared to the model utilizing the Distance feature, XGBoost is the model to implement for predictions on new unseen data regardless of distance metric used.

4.1.9. Feature reduction and re-evaluation

In previous chapters, the models were trained and evaluated, while 19 features were used in total, including those that were one hot encoded. The models were re-evaluated after dropping most of the features. By observing the plots of the model importance, the 7 most influential features for tree construction were selected. Those features are **Income, Time, Convenience, Cost, Health, Weather and Distance**. The new models were initially trained on the training set. Afterwards, a grid search was executed to identify the new optimal parameter values and the models were evaluated on the test set afterwards.

A) Decision tree

For decision tree, a grid was run with the following parameters:

```
param_grid = {
   'max_depth': [2, 3, 4, 5, 6],
```

'min_samples_split': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],

'min_samples_leaf': [1, 2, 3, 4, 5, 6, 7]

The best results were for **max_depth = 6, min_samples_leaf=1 and min_samples_split=3 with a validation accuracy of 0.868**. The table below illustrates the performance of the new tree on the test set compared to the one using all the 19 features.

Table 26. Pruned Tree (7 vs 19 features)

	Tree (7 features)			Tree (19 features)		
	Private vehicle	Bus	Walk	Private vehicle	Bus	Walk
Precision	0.81	0.91	0.78	0.82	0.89	0.80
Recall	0.85	0.82	0.82	0.82	0.80	0.88
F1-score	0.83	0.86	0.80	0.82	0.84	0.83
Accuracy	0.83			0.83		
AUC-Macro	0.88			0.90		

It is observed that feature reduction had minor changes on the metrics for the decision tree. Recall for bus increased from 0.80 to 0.82, while for private vehicle also increased from 0.82 to 0.85. In contrast, recall for walk dropped from 0.88 to 0.82. Precision only increased for bus label, from 0.89 to 0.91, while it slightly dropped for private vehicle and walk. AUC macro score is slightly higher for the tree utilizing all of 19 features, demonstrating slightly better performance.

The same procedure was executed by swapping Distance with geodesic distance. A new grid was run to find the optimal values. The results were best for **max_depth = 5, min_sampels_leaf=1, min_samples_split=2 and a validation accuracy of 0.845.** The table below shows the results of the new tree compared to the tree utilizing all of 19 features.

Table 27. Pruned Tree - Geodesic Distance (7 vs 19 features)

	Geodesic Tree (7 features)			Geodesic Tree (19 features)		
	Private vehicle	Bus	Walk	Private vehicle	Bus	Walk
Precision	0.79	0.87	0.81	0.81	0.78	0.88
Recall	0.96	0.69	0.78	0.87	0.82	0.78
F1-score	0.87	0.77	0.79	0.84	0.80	0.83
Accuracy	0.82			0.82		
AUC-Macro	0.89			0.88		

For private vehicle class, recall increased from 0.87 to 0.96. In contrast, recall for bus dropped from 0.82 to 0.69. Recall for walk remained the same at 0.78. Precision increased for bus, from 0.78 to 0.87 while it dropped for walk and private vehicle at 0.81 and 0.79 respectively. The overall accuracy remained the same at 0.82. While the auc macro is slightly higher at 0.89. Overall, feature reduction did not significantly improve the performance of the trees regardless of distance metric.

B) ***Random Forest***

For Random Forest the same parameters were examined as before: the number of trees, max features, and max tree depth. A grid search was run for each parameter individually. The results are illustrated on the plots below.

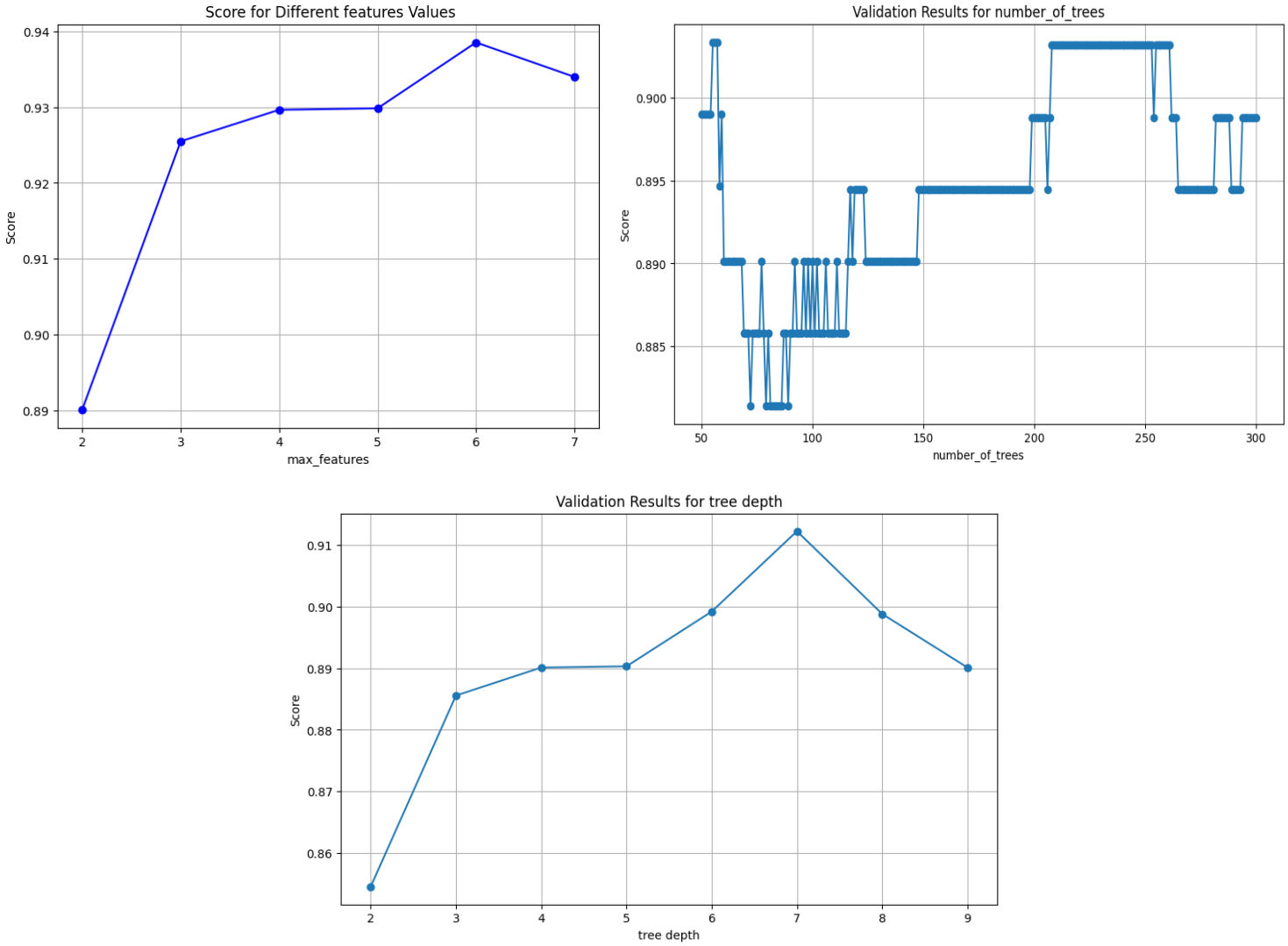


Figure 83. Grid Search RF (7 features)

After finding the optimal area, a final grid was run using the following parameter values.

```
param_grid = {
    'n_estimators': list(range(50, 61)),
    'max_features': [6],
    'max_depth' : [7]}
```

The best results are for **max depth =7, max features =6, number of trees = 55 and a validation accuracy of 0.934**. The model was then evaluated on the test set. The table below depicts the differences of the two Random Forest models (7 vs 19 features).

Table 28. Random Forest (7 vs 19 features)

	Random Forest (7 features)			Random Forest (19 features)		
	Private vehicle	Bus	Walk	Private vehicle	Bus	Walk
Precision	0.96	0.98	0.87	0.96	0.95	0.84
Recall	0.93	0.92	0.96	0.93	0.86	0.96
F1-score	0.94	0.95	0.91	0.94	0.90	0.90
Accuracy	0.93			0.92		
AUC-Macro	0.985			0.983		

For the new model, all metrics for private vehicle remained the same. Both recall and precision increased at 0.92 and 0.98 respectively. Precision also increased for walk, from 0.84 to 0.87 while recall remained the same at 0.96. The overall accuracy increased from 0.92 to 0.93 while auc macro slightly improved from 0.983 to 0.985. Overall, by reducing the features of the model, increased the performance of Random Forest.

Distance was again swapped with Geodesic distance. The new optimal parameters are illustrated on the plots below.

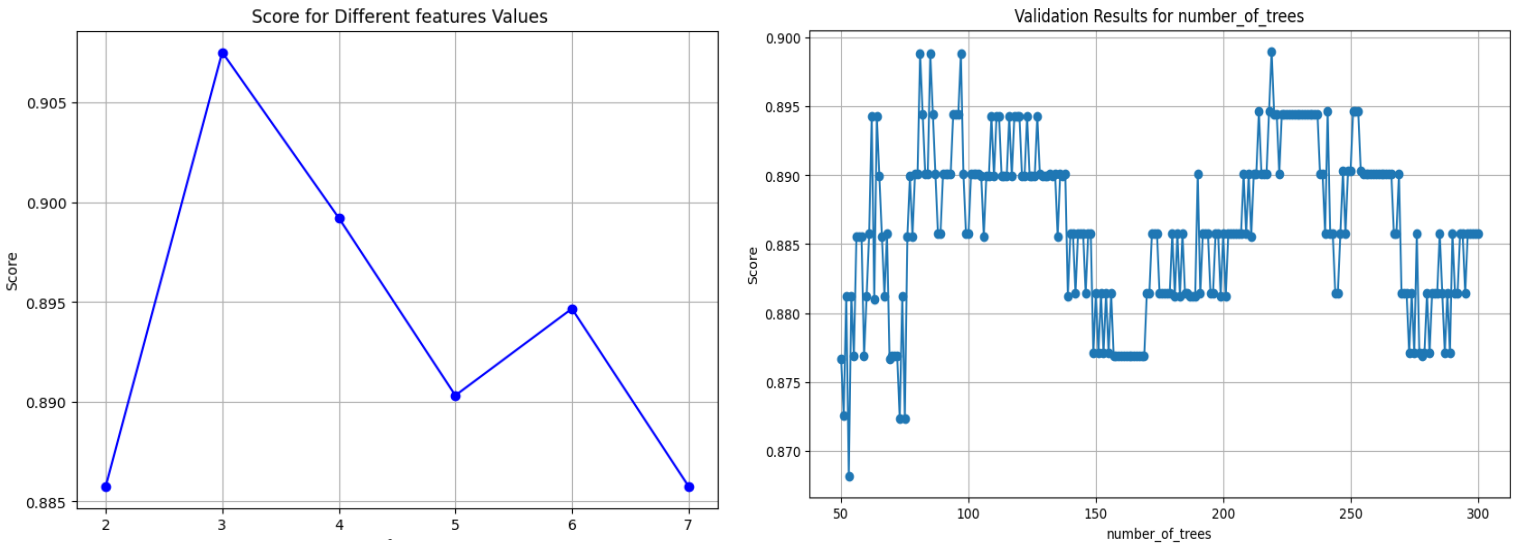


Figure 84. Grid Search RF - 7 features

After finding the optimal area, a new grid search was run with the following parameter values.

param_grid = {

'n_estimators': list(range(60, 101)),

'max_features': [3]}

The best results are for max features = 3, number of trees = 68 and a validation accuracy of 0.91. The model was then evaluated on the test set. The results are illustrated below compared to the model using all 19 features.

Table 29. Random Forest – Geodesic Distance (7 vs 19 features)

	Geodesic Random Forest (7 features)			Geodesic Random Forest (19 features)		
	Private vehicle	Bus	Walk	Private vehicle	Bus	Walk
Precision	0.90	0.89	0.84	0.88	0.95	0.91
Recall	0.95	0.84	0.84	0.88	0.78	0.83
F1-score	0.92	0.86	0.84	0.80	0.84	0.82
Accuracy	0.88			0.86		
AUC-Macro	0.96			0.96		

For private vehicle, recall increased from 0.88 to 0.95. For walk it also increased from 0.83 to 0.84, while for bus it also increased from 0.78 to 0.84. Precision increased for private vehicle from 0.88 to 0.90 while it dropped for the other two classes. The overall accuracy of the model increased from 0.86 to 0.88 while the auc macro score remained approximately the same. By reducing the features, the Random Forest model (using geodesic distance) demonstrates a more balanced performance, also retrieving more relevant instances for each class.

C) *XGBoost*

For XGBoost, the same parameters were tuned as before. The values that were configured in grid search are depicted below.

```
param_grid = {
    'n_estimators': list(range(100, 151)),
    'subsample': [0.3],
    'colsample_bytree' : [0.9],
    'learning_rate': [0.2],
    'max_depth':[2]}
```

The best result was for **number of trees =100, max depth=2, learning rate=0.2, subsample=0.3, colsample =0.9 with a validation accuracy of 0.947**. The comparisons of the two models (7 vs 19 features) on the test set are illustrated below.

Table 30. XGBoost (7 vs 19 features)

	XGBoost (7 features)			XGBoost (19 features)		
	Private vehicle	Bus	Walk	Private vehicle	Bus	Walk
Precision	0.93	0.96	0.94	0.95	0.96	0.88
Recall	0.98	0.92	0.92	0.96	0.88	0.94
F1-score	0.96	0.94	0.93	0.95	0.91	0.91
Accuracy	0.94			0.93		
AUC-Macro	0.99			0.99		

Recall for private vehicle increased from 0.96 to 0.98 with a trade-off in precision as it dropped at 0.93. Recall also increased for bus from 0.88 to 0.92, while precision remained the same at 0.96. Recall for walk dropped from 0.94 to 0.92 while precision increased from 0.88 to 0.94. The overall accuracy of the model increased from 0.93 to 0.94 while the auc macro remained approximately the same at 0.99. By reducing the features, the model demonstrates a more balanced performance across the classes retrieving also more relevant instances.

Again, the model was evaluated swapping distance with geodesic distance. The new parameter values that were tuned through grid search are depicted below.

```
param_grid = {
    'n_estimators': list(range(50, 61)),
    'subsample': [0.3],
    'colsample_bytree' : [0.8],
    'learning_rate': [0.3],
    'max_depth':[2]}
```

The best parameters were found for **max depth =2, learning rate =0.3, number of trees =50, subsample=0.3, colsample=0.8 and a validation accuracy of 0.89**.
The table below illustrates the differences in the two models (7 features vs 19 features).

Table 31. XGBoost - Geodesic Distance (7 vs 19 features)

	Geodesic XGBoost (7 features)			Geodesic XGBoost (19 features)		
	Private vehicle	Bus	Walk	Private vehicle	Bus	Walk
Precision	0.88	0.88	0.85	0.87	0.91	0.85
Recall	0.91	0.86	0.84	0.98	0.84	0.80
F1-score	0.89	0.84	0.85	0.92	0.87	0.82
Accuracy	0.87			0.88		
AUC-Macro	0.969			0.967		

Recall for private vehicle decreased from 0.98 to 0.91 with a slight increase in precision, from 0.87 to 0.88. Recall for bus increased from 0.94 to 0.96. Recall for walk also increased from 0.80 to 0.84. The overall accuracy slightly drooped from 0.88 to 0.87, while auc macro slightly increased from 0.967 to 0.969. Overall, there is not a significant difference in the model by reducing the number of features.

D) ***Stacked Model***

Finally, new stacked models were constructd using only the 7 available features. The tuned models were used as base estimators (Decision Tree, Random Forest, XGBoost), while Logistic Regression (one vs rest) was used as the meta learner. The table below illustrates the differences in the models utilizing Distance as a feature.

Table 32. Stacked Model (7 vs 19 features)

	Stacked (7 features)			Stacked (19 features)		
	Private vehicle	Bus	Walk	Private vehicle	Bus	Walk
Precision	0.94	0.86	0.94	0.96	0.86	0.86
Recall	0.91	0.90	0.94	0.93	0.86	0.90
F1-score	0.93	0.88	0.94	0.94	0.86	0.88
Accuracy	0.92			0.90		
AUC-Macro	0.958			0.93		

Recall increased for bus, from 0.86 to 0.90 while precision remained at 0.86. Recall for walk also increased from 0.90 to 0.94, while precision increased from 0.86 to 0.94. In contrast, recall for private vehicle decreased from 0.93 to 0.91. The overall accuracy of the model increased from 0.90 to 0.92, while auc macro also increased from 0.93 to 0.958. Overall, by reducing the features the model demonstrates a more balanced performance for the individual classes.

Distance metrics were again swapped and evaluate the new stacked model with reduced number of features. The tuned tree (Decision Tree, Random Forest and XGBoost) using geodesic distance, were used as the base estimators, while Logistic Regression was used as the meta learner. The comparison of the models is illustrated below (Geodesic distance – 7 features vs 19 features).

Table 33. Stacked Model - Geodesic Distance (7 vs 19 features)

	Geodesic Stacked (7 features)			Geodesic Stacked (19 features)		
	Private vehicle	Bus	Walk	Private vehicle	Bus	Walk
Precision	0.86	0.86	0.79	0.87	0.87	0.85
Recall	0.93	0.73	0.84	0.96	0.80	0.82
F1-score	0.89	0.79	0.81	0.91	0.83	0.83
Accuracy	0.84			0.86		
AUC-Macro	0.95			0.926		

Recall for private vehicle reduced from 0.96 to 0.93. Recall for bus also dropped from 0.80 to 0.73. In contrast, recall for walk increased from 0.82 to 0.84. Precision for all individual classes also dropped. The overall accuracy of the model dropped from 0.86 to 0.84. By reducing the features and using geodesic distance, stacked model seems to perform worse compared to the original, since it retrieves lower number of relevant instances, while precision is lower for all classes.

The roc curves were constructed for the new models (using 7 features). The results, along with the model comparison table are illustrated below, initially for the models utilizing Distance metric.

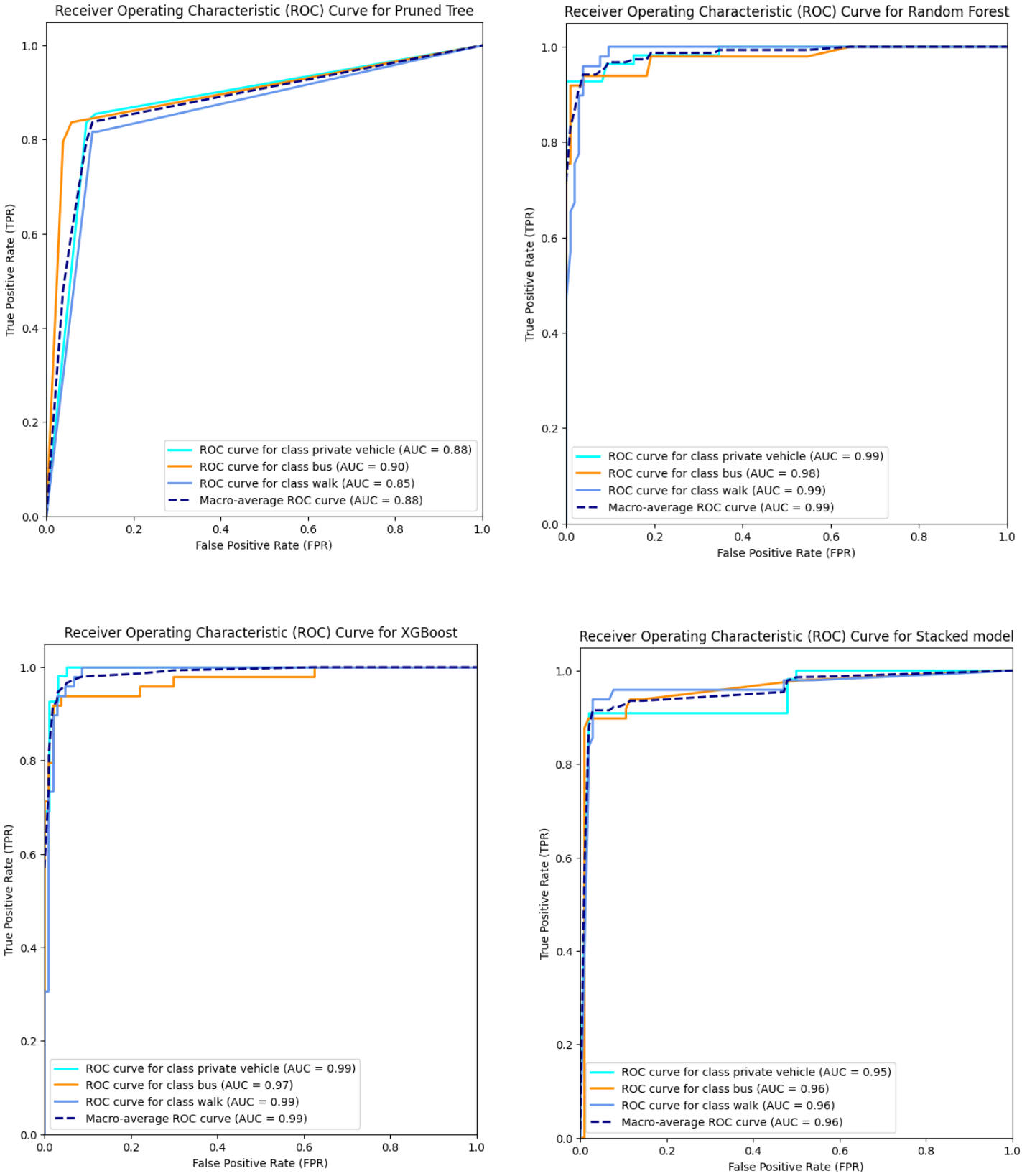


Figure 85. ROC Curves - 7 features

Table 34. Model comparison - 7 features

Models with Distance and Reduced Features	Decision Tree	Random Forest	XGBoost	Stacked Model
Precision (Average)	0.83	0.94	**0.94**	0.91
Recall (Average)	0.83	0.93	**0.94**	0.92
F-Score (Average)	0.83	0.93	**0.94**	0.91
AUC (Average)	0.876	0.985	**0.988**	0.958
Accuracy	0.83	0.93	**0.94**	0.92

The XGBoost model stands out as the top-performing model across precision, recall, and F-score, demonstrating consistently high values. Random Forest closely follows, exhibiting high precision and recall scores, and a high AUC value of 0.985. While Stacked Model also displays commendable performance, XGBoost and Random Forest emerge as particularly better choices for predicting transportation mode utilizing Distance and a reduced set of features. Overall, these models showcase strong predictive capabilities, with XGBoost and Random Forest leading the way in terms of comprehensive performance metrics.

The same procedure was followed for the models utilizing Geodesic distance. The results are illustrated below.

Table 35. Model comparison - 7 features

Models with Geodesic Distance and Reduced Features	Decision Tree	Random Forest	XGBoost	Stacked Model
Precision (Average)	0.82	**0.87**	**0.87**	0.84
Recall (Average)	0.81	**0.87**	**0.87**	0.83
F-Score (Average)	0.81	**0.87**	**0.87**	0.83
AUC (Average)	0.89	**0.96**	**0.96**	0.93
Accuracy	0.82	**0.88**	**0.87**	0.84

Random Forest continues to showcase increased performance, though lower when compared to the previous models. XGBoost closely follows, maintaining a balanced performance across various metrics. Decision Tree and the Stacked Model demonstrate mediocre results, albeit slightly lower than those achieved with distance and reduced features. Even by reducing the number of feature, using geodesic distance did not improve the performance of the classifiers significantly.

4.1.10. Model selection.

The table below illustrates the final performance measures for the best performing models across all the previous chapters for a final evaluation.

Table 36. Final Model comparison

	Precision (Average)	Recall (Average)	F1 Score (Average)	Accuracy
Decision Tree (19 features)	0.83	0.83	0.83	0.83
Decision Tree (7 features)	0.83	0.83	0.83	0.83
Decision Tree (Geodesic-19 features)	0.83	0.82	0.82	0.82
Decision Tree (Geodesic – 7 features)	0.82	0.81	0.81	0.82
Random Forest (19 features)	0.92	0.91	0.91	0.92
Random Forest (7 features)	0.94	0.93	0.93	0.93
Random Forest (Geodesic-19 features)	0.86	0.86	0.85	0.86
Random Forest (Geodesic – 7 features)	0.87	0.87	0.87	0.88
XGBoost (19 features)	0.93	0.93	0.93	0.93
XGBoost (7 features)	**0.94**	**0.94**	**0.94**	**0.94**

XGBoost (Geodesic-19 features)	0.88	0.87	0.87	0.88
XGBoost (Geodesic – 7 features)	0.87	0.87	0.87	0.87
Stacked (19 features)	0.89	0.89	0.89	0.90
Stacked (7 features)	0.91	0.92	0.91	0.92
Stacked (Geodesic-19 features)	0.86	0.86	0.86	0.86
Stacked (Geodesic – 7 features)	0.84	0.83	0.83	0.84

Decision Tree Models:

- For Decision Tree models with both 19 features and 7 features, there is consistency in performance across precision, recall, F1 score, and accuracy, all achieving a score of around 0.83. This suggests that the model's predictive ability remains stable even with a reduced feature set. Though the models are prone to overfitting since there is a huge gap between training and testing performances.
- Introducing geodesic distance has a slightly different impact. While precision remains similar, there is a slight decrease in recall, F1 score, and accuracy. This may indicate that the geodesic distance introduces some complexity making difficult for trees to classify effectively and hindering predictive performance.

Random Forest Models:

- Random Forest models consistently outperform Decision Trees. The model with 19 features achieves high precision (0.92) and recall (0.91), leading to an increased F1 score (0.91) and accuracy (0.92). By reducing the features, the performance of the model increases, achieving high precision (0.94), recall (0.93), F1 score (0.93) and an increased overall accuracy (0.93).

- Like Decision Trees, introducing geodesic distance in Random Forest models leads to a decrease in performance. The model with 19 geodesic features shows a noticeable drop in precision, recall, F1 score, and accuracy.

XGBoost Models:

- XGBoost models exhibit strong performance across all metrics, with the model using 7 features achieving particularly high scores (precision: 0.94, recall: 0.94, F1 score: 0.94, accuracy: 0.94). This highlights the effectiveness of XGBoost for the specific dataset.
- Like Decision Trees and Random Forests, introducing geodesic distance has a marginal negative impact on performance.

Stacked Models:

- Stacked models, which combine predictions from multiple base models, demonstrate competitive performance. The model with 7 features achieves high precision (0.91), recall (0.92), F1 score (0.91), and accuracy (0.92). However, they fall behind compared to the individual Random Forest and XGBoost models.
- Geodesic Distance again led to a decrease in performance across all metrics for stacked models.

Overall Observations:

- Random Forest and XGBoost consistently outperform Decision Trees and Stacked models in terms of precision, recall, F1 score, and accuracy.
- The impact of geodesic distance varies across models, with a general trend of a slight decrease in performance. This suggests that while geodesic distance is easier to compute using the geopy library, there is an increase in complexity that hinders predictive performance in these specific scenarios.
- The choice of the number of features significantly influences model performance. Models with fewer features (7 features) often achieve higher precision, recall, F1 score, and accuracy compared to models with a larger number of features (19 features).
- The best performing model is XGBoost utilizing distance and 7 features in total.

4.1.11. Model Explainer

A fundamental challenge in machine learning models lies in comprehending feature importance and interpreting the outcomes. Although tree-based models visualize this importance through feature importance functions, these essentially reflect how the models, and thus individual trees, were constructed. Referring to the feature importance of Random Forest models, it is evident that distance and time consistently hold the highest positions in the list of influential features, irrespective of the number of features considered. This is illustrated in the plots below.

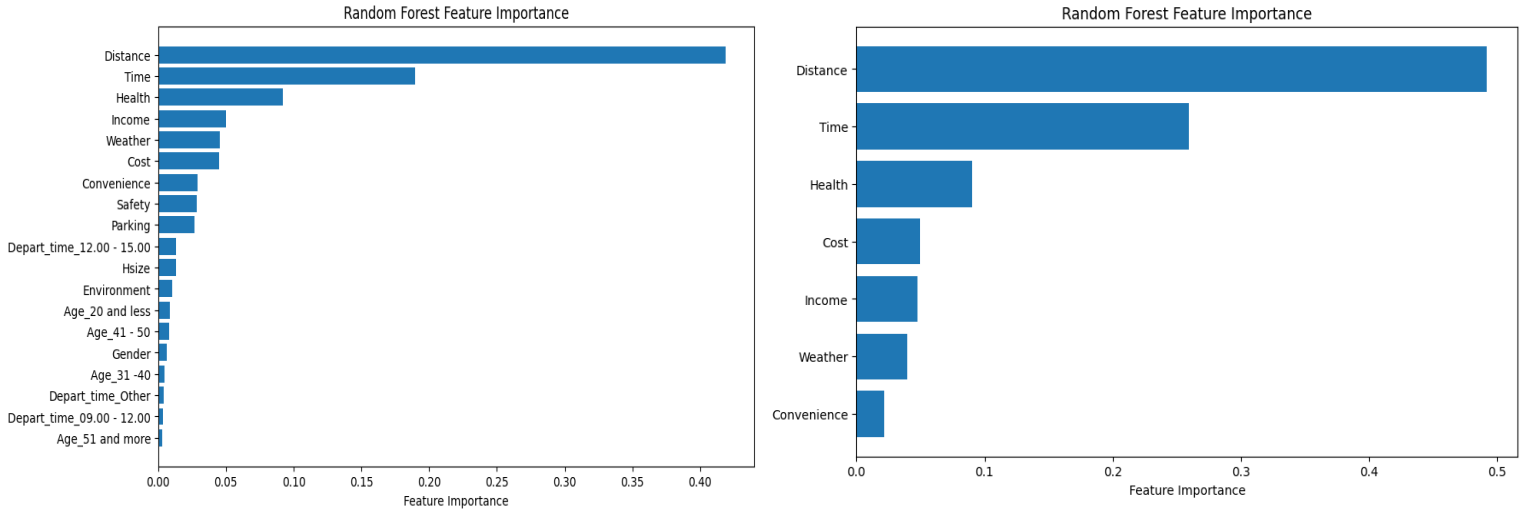


Figure 86. Feature Importance RF (7 vs 19 features)

However, the outcomes become less straightforward with the XGBoost model. While distance maintains its prominence as the most crucial feature, it appears that most of the features contribute significantly to the construction of individual trees. This is particularly evident in the model employing seven features, where factors like physical activity and health take precedence over time in terms of importance. The subsequent features also demonstrate notable contributions. These findings are visually represented below.

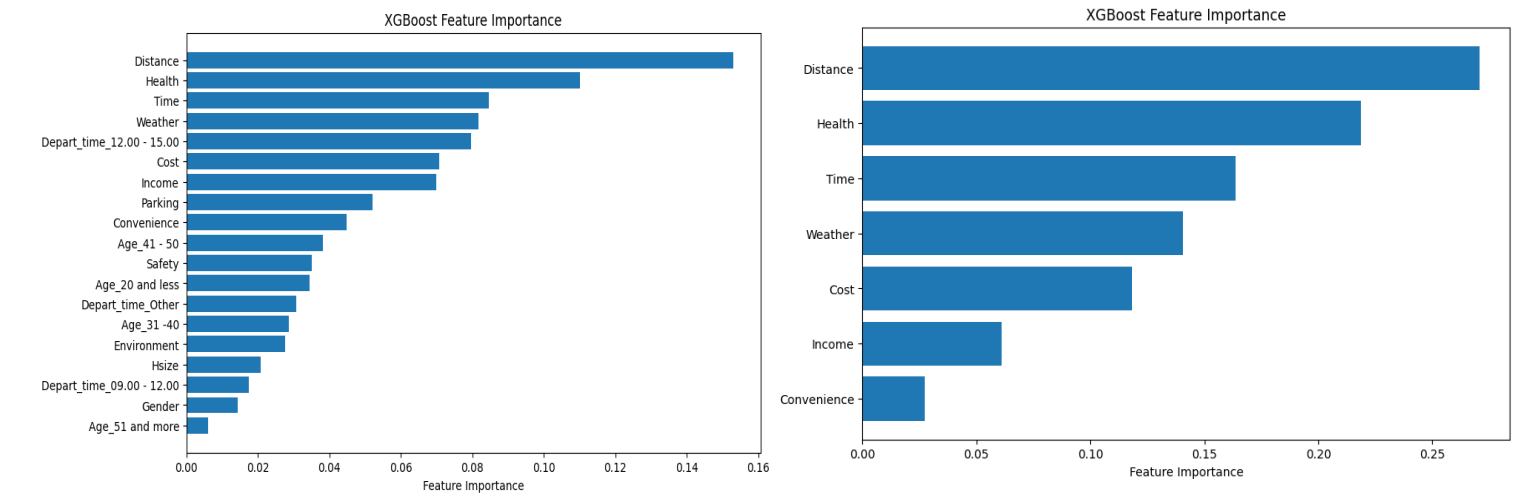
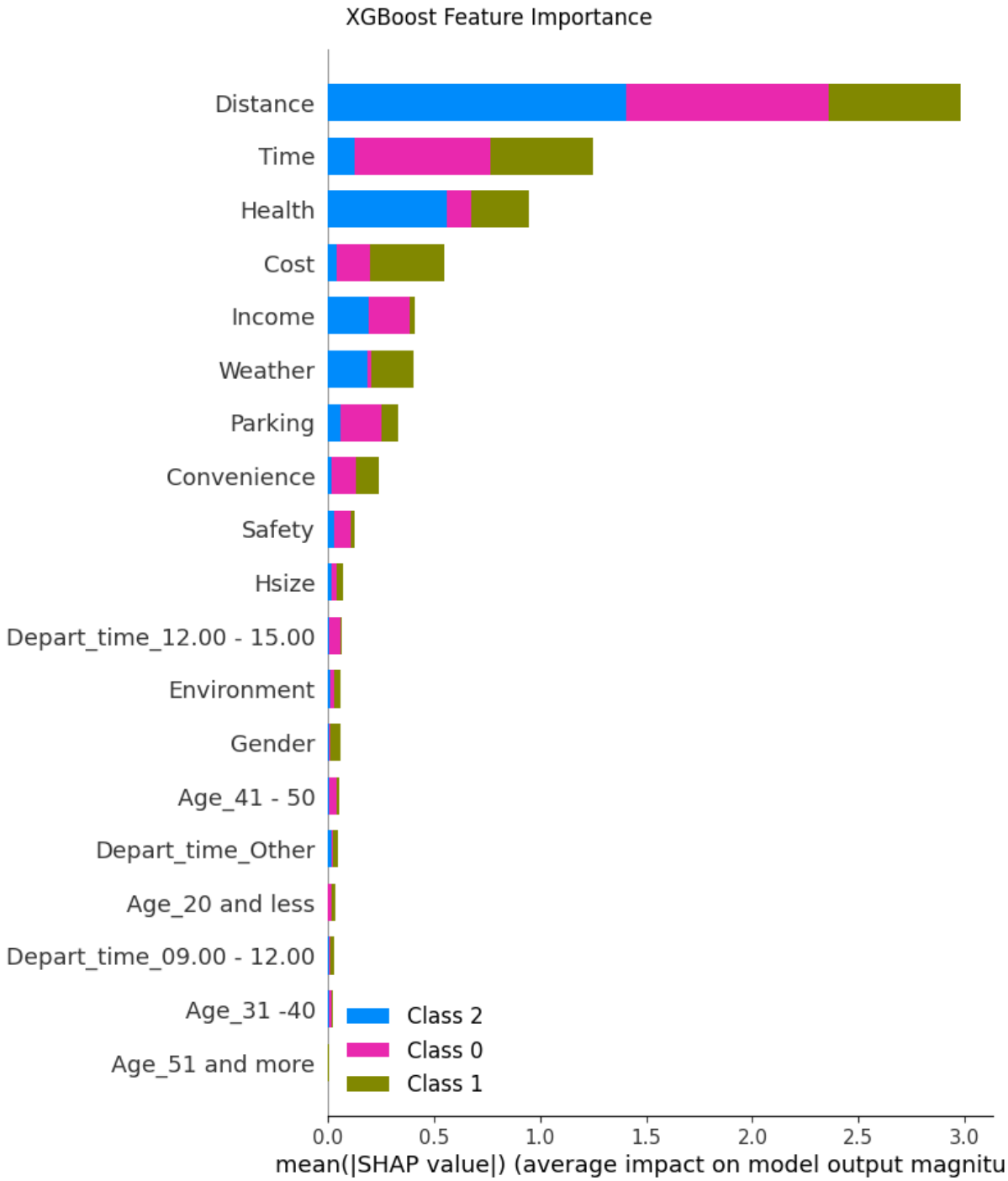


Figure 87. XGB feature importance (7 vs 19 features)

To address this question, the shap Python library comes into play. The library is constructed based on the "SHAP (Shapley Additive exPlanation) theory," as proposed by Lundberg & Lee (2017). SHAP is a method designed for interpreting the output of machine learning models. It introduces a framework to explain a model's predictions by attributing the prediction to each individual feature. The SHAP values that the library computes, provide insights into how each feature influences individual predictions, highlighting the significance of each feature relative to others and showcasing the model's dependence on feature interactions. They offer a consistent and objective method for explaining the impact of each feature on the model's predictions. Based on cooperative game theory, SHAP values assign importance values to each feature. Assume a cooperative game with players equal to the number of features. In this scenario, SHAP reveals the individual contribution of each "player" (or feature in machine learning) to the model's output for every example or observation (Trevisan, 2022). Features with positive SHAP values positively influence predictions, while those with negative values exert a negative impact. The magnitude of the SHAP values indicates the strength of the effect (Awan, 2023).
The shap explainer function was used to construct the feature importance for XGBoost model (19 features). The code is depicted below. Note that ‘best_model’ refers to the tuned model with the best parameters found through grid search.

```
explainer = shap.Explainer(best_model)
shap_values = explainer.shap_values(X_test)
```

The summary plot was then created to display feature importance. The plot displays the mean SHAP values of each feature and for each class label.



Class 0: Private Vehicle, Class 1: Bus, Class 2: Walk

Figure 88 XGBoost feature importance - 19 features - Shap

From the plot it is identified that Distance was indeed the most important feature for predictions regardless of class label. Time was the second most important feature, while the magnitude of the importance was higher for private vehicle – bus and lower for walk label. In contrast, the magnitude of

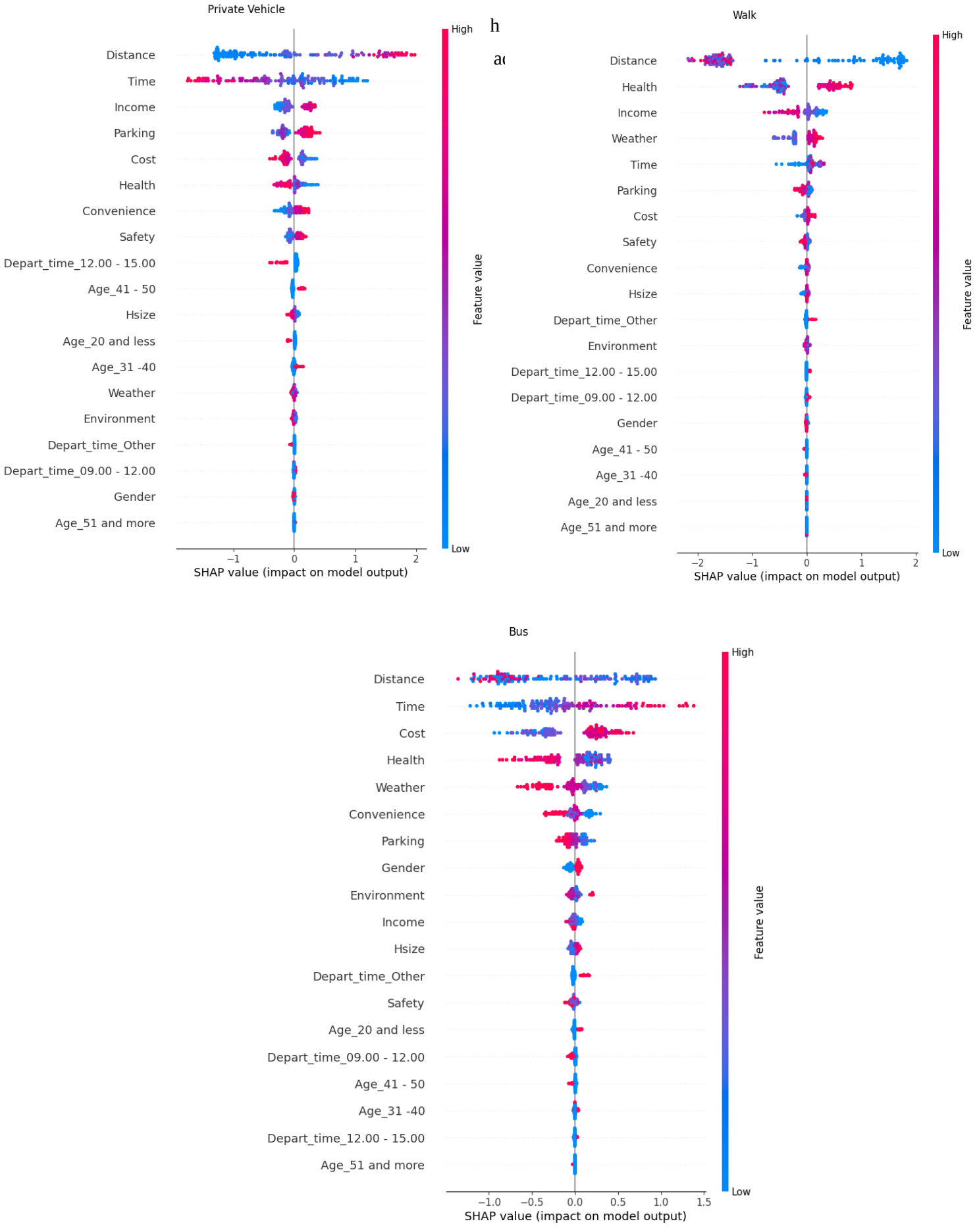


Figure 89. XGB explainer - 19 features

The above plots illustrate the impact of each feature for each class on the whole dataset. Each scatter represents an observation of the test set. Each colour describes the feature value (red for high and blue for low) while the values of the X axis indicate the impact on the outcome. The features are put in descending order from the most important feature to the least important one. A plot is generated for each class respectively. To interpret the plot, focus should be given at a specific class. For instance, by observing the plot of Walk class, it is identified that distance was the most important feature, followed by health. Low values (blue colour) for Distance (referring to short travel distances in kilometres) have a high positive contribution for predicting that class. Additionally, high values (red colour) for Health (referring to Agree and Strongly Agree on the impact of physical activities and health) have also a positive contribution on predicting that class, though the magnitude is lower that of distance. In contrast, high values (red colour) for Income (referring to commuters with income above 1000) have a negative contribution on predicting that class. In a multiclassification problem the negative contribution implies not predicted as that class (Walk vs rest of modes). For a more generic interpretation, the shorter the distances for commuting to work, the higher the tendency of the model to predict walking. In contrast, the highest the income, the lower the tendency of the model to predict walk.

Note that while Health is the second most important feature for predicting walk class, this is not the same for private vehicle and bus modes. By observing the plot for Bus, Distance is first followed by Time. The interpretation is the same as before. Higher values (red colour) for Time (referring to the minutes required for commuting to work) have a high positive impact in predicting that class. The generic interpretation is still, the higher the commute time to work, the higher the tendency of predicting bus. In contrast, higher values (red colour) for Health (Agree, strongly Agree), have a negative impact in predicting that class. Hence, as the ranking position increases, from agree to strongly agree, the tendency of predicting bus decreases (higher chance that the commute happens with other mode and possibly walking). The same pattern could be followed for every feature and for every class in trying to explain the outcome of the prediction. The same graphs were computed for XGBoost model utilizing 7 features. The summary plot of feature importance is illustrated below.

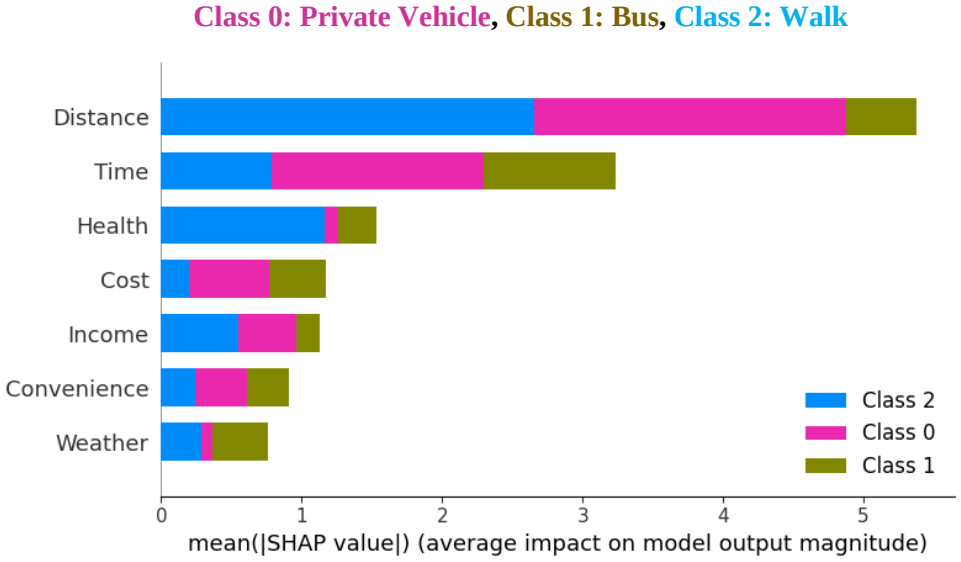


Figure 90. XGB feature importance - 7 features

Once more, Distance emerges as the paramount feature but only for private vehicle and walk now, compared to all classes in the previous model. For bus, Time has surpassed Distance as the most crucial feature, while Health is the second most important for walk labels, as indicated by the magnitude of the values. The three individual plots were constructed for a better understanding of the feature importance on predictions.

Figure 91. XGB Explainer - 7 features

By observing the plot for Bus, the interpretation is the same as before. Higher values (red colour) for Time (referring to the minutes required for commuting to work) have a high positive impact in predicting that class. The generic interpretation is still, the higher the commute time to work, the higher the tendency of predicting bus. In contrast, for private vehicle, the higher the distance, the higher the tendency of predicting private vehicle. The same procedure can be followed for every iteration of class and features just by observing the individual plots.

In general, SHAP analysis offers a more detailed and individualized approach to feature importance compared to traditional methods like retrieving feature importance for a tree-based model in scikit-learn. While scikit-learn provides feature importance based on metrics like Gini impurity or information gain, it tends to offer a global ranking without considering the interaction effects between features. SHAP, on the other hand, provides a specialised approach that considers the contribution of each feature for every individual prediction. This allows for a better understanding of how each feature influences specific instances, capturing both main effects and interactions. Shap library contains additional plots that could be investigated, such as dependency plots, while it is possible to explore the magnitude of each prediction of the model. Although this kind of analysis is not explored is this Thesis, it could very well be a fundamental context for research in another work.

5. CONCLUSIONS

5.1. Summary

In this chapter, the Thesis summarizes its findings, commencing with an overview of popular concepts, models, and evaluation metrics in machine learning derived from an extensive literature review. Notably, tree-based models emerge as the prevailing techniques for both classification and regression tasks, exhibiting heightened accuracy compared to traditional logit models, particularly in the realm of transport mode choice. The central focus of the Thesis lies in predicting the transport mode choice of residents in Thessaloniki, specifically during their commutes to work. Data was collected through surveys, both online and in print, from active workers in the city, resulting in a total of 409 samples. After preparation and cleaning, 381 samples were deemed suitable for further analysis. The dataset comprises various trip-related features, such as distance, commute and departure time, alongside demographic characteristics like age, income, and household size. Ordinal scale features, like cost or physical activity impact on commutes, were also included on a scale of 1 to 5. Following an exploratory data analysis, preprocessing steps were implemented to make the data applicable for modelling in the Python programming language. These steps encompassed encoding, train/test splitting, and normalization of data within a 0-1 range.

Four models were employed in total: Decision Tree, Random Forest, XGBoost, and a Stacked model combining the previous three. Initial training involved 19 features, and various iterations were explored, including parameter tuning and distance metric swapping. The first experimentation, utilizing all features, revealed that Random Forest and XGBoost models exhibited the best performance, achieving overall accuracies of 0.92 and 0.93, respectively. However, substituting distance with the geodesic distance metric significantly diminished classifier performance. The most critical features for the models were identified as Distance, Duration, Health, Cost, Income, and Weather. As a secondary experiment, new models were constructed using only the most crucial features (7 in total). Excluding the Decision Tree, all models demonstrated improved performance, with Random Forest and XGBoost achieving overall accuracies of 0.93 and 0.94, respectively. Distance and time remained paramount among the features for these models.

Moreover, the SHAP explainer library was employed to further investigate feature importance for each class individually, providing a visualization of the model's decision-making process. Notably, XGBoost, being the best-performing model, identified distance and time as the primary influencers for the private vehicle class, while other classes exhibited minor distinctions. For the bus class, time emerged as the most crucial factor, whereas for the walk class, distance took precedence, followed by physical activity and health. The SHAP library proved to be a valuable tool in enhancing the understanding of how the model operates and what factors influence its predictions.

5.2. Limitations and Future work

In this section of the Thesis, limitations and future research ideas are provided. Geographical specificity poses a primary constraint on the applicability of the developed models. Tailored to the unique commuting behaviors of Thessaloniki residents, these models may lack generalizability to other cities or regions due to variations in infrastructure, cultural preferences, and commuting norms. The variations of different locations may require the development of city-specific models for accurate predictions. Moreover, the models are currently optimized for the classification of private vehicle, bus, and walk modes. To extend their applicability to additional modes such as bikes or skates, a broader and more diverse dataset would be essential.

Temporal dynamics represent another limitation, as the dataset used for training and evaluation captures commuting behavior within a specific timeframe. The models may not adequately account for shifts in infrastructure, economic conditions, or lifestyle choices over time, impacting their predictive capabilities. Consideration of temporal factors is crucial for maintaining the models' relevance and accuracy. Moreover, the models assume a relatively stable environment without accounting for potential mode shifts or interventions. Changes in transportation modes or the introduction of new public transport initiatives are not considered. Future work should explore ways to integrate these dynamic elements into the models, ensuring adaptability to sudden shifts or external interventions affecting commuting patterns. The reliance on survey data introduces biases inherent in self-reporting. Respondents may not always provide accurate information regarding their commuting behaviours or preferences. Additionally, the representativeness of the sample size may be limited, potentially leading to biases in the predictions generated by the models.

Future research endeavours can explore cross-city generalization by developing models that transcend specific locations. Collecting diverse datasets from multiple cities or regions can facilitate the creation of models adaptable to different commuting contexts. This approach aligns with the goal of creating versatile models capable of addressing the unique features of various urban environments.

Incorporating real-time data sources, such as traffic updates, weather conditions, and public transport schedules, stands as a key improvement for future models. This integration enhances the accuracy and responsiveness of models to immediate changes in commuting environments, providing users with more reliable predictions. The inclusion of behavioural factors, such as attitudes towards sustainable transportation or the influence of social networks, can contribute to a more comprehensive understanding of mode choice decisions. Future models can explore the impact of psychological and social factors on commuting behaviour, providing valuable insights into the decision-making process. Additionally, future research could focus on evaluating the impact of transportation interventions or policy changes on commuting behaviours. Understanding how models respond to significant external interventions enhances their utility in assessing the effectiveness of urban planning strategies and policy implementations.

Lastly, improving user interaction and model explainability should be a priority for future developments. Creating user-friendly interfaces and visualizations that facilitate the interpretation of model predictions enhances user trust and engagement, contributing to the broader applicability of the models.

REFERENCES

Ababio-Donkor, A., Saleh, W. & Fonzone, A. 2020, "Understanding transport mode choice for commuting: the role of affect", *Transportation planning and technology*, vol. 43, no. 4, pp. 385-403.

Amiri, I.S., Akanbi, O.A. and Fazeldehkordi, E. (2014) *A Machine-Learning Approach to Phishing Detection and Defense*. 1st edn. Rockland, MA: Elsevier Science.

Amor, N.B., Benferhat, S. and Elouedi, Z. (2006) "Qualitative Classification with Possibilistic Decision Trees," in *Modern Information Processing*. Elsevier B.V, pp. 159–169.

Awan, A.A. (2023) *An introduction to shap values and machine learning interpretability*, *DataCamp*. Available at: https://www.datacamp.com/tutorial/introduction-to-shap-values-machine-learning-interpretability (Accessed: 27 February 2024).

Bei, H. et al. (2023) "Joint prediction of travel mode choice and purpose from travel surveys: A multitask deep learning approach," Travel, behaviour & society, 33, p. 100625.

Bellini, T. (2019) *IFRS 9 and CECL credit risk modelling and validation a practical guide with examples worked in R and SAS*. London: Academic Press.

Bhuiya, M.M.R. et al. (2022) "Application of Machine Learning Classifiers for Mode Choice Modeling for Movement-Challenged Persons," Future transportation, 2(2), pp. 328–346.

Böcker, L., Dijst, M. and Faber, J. (2016) "Weather, transport mode choices and emotional travel experiences," Transportation research. Part A, Policy and practice, 94, pp. 360–373.

Brownlee, J. (2020) *ROC curves and precision-recall curves for imbalanced classification*, *MachineLearningMastery.com*. Available at: https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-imbalanced-classification/ (Accessed: 20 February 2024).

Brownlee, J. (2021) Random oversampling and undersampling for imbalanced classification, MachineLearningMastery.com. Available at: https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/ (Accessed: 12 January 2024).

Brownlee, J. (2021) *Stacking Ensemble Machine Learning with python*, *MachineLearningMastery.com*. Available at: https://machinelearningmastery.com/stacking-ensemble-machine-learning-with-python/ (Accessed: 20 February 2024).

Cunningham, P. and Delany, S.J. (2021) "k-Nearest Neighbour Classifiers - A Tutorial," *ACM computing surveys*, 54(6), pp. 1–25.

Das, S. *et al.* (2021) "Impact of COVID-19: A radical modal shift from public to private transport mode," *Transport policy*, 109, pp. 1–11.

Hagenauer, J. and Helbich, M. (2017) “A comparative study of machine learning classifiers for modeling travel mode choice,” Expert systems with applications, 78, pp. 273–282.

Kettle, S. (2014) Geodesic distances: How long is that line again?, Esri Community. Available at: https://community.esri.com/t5/coordinate-reference-systems-blog/geodesic-distances-how-long-is-that-line-again/ba-p/902188 (Accessed: 25 January 2024).

Kubat, M. (2021) An Introduction to Machine Learning. 3rd edn. Cham: Springer International Publishing AG.

Kufel, J. et al. (2023) “What Is Machine Learning, Artificial Neural Networks and Deep Learning?-Examples of Practical Applications in Medicine,” Diagnostics (Basel), 13(15), p. 2582.

Lee, E. and Kim, S. (2023) *Geographic Information Systems for Intermodal Transportation*. Elsevier.

Lopez, F. (2021) Ensemble learning: Bagging & boosting, Medium. Available at: https://towardsdatascience.com/ensemble-learning-bagging-boosting-3098079e5422 (Accessed: 12 January 2024).

Lundberg, S. and Lee, S.-I. (2017) ‘A Unified Approach to Interpreting Model Predictions’, *arXiv.org* [Preprint]. Available at: https://doi.org/10.48550/arxiv.1705.07874.

Mayo, F.L. & Taboada, E.B. 2020, "Ranking factors affecting public transport mode choice of commuters in an urban city of a developing country using analytic hierarchy process: The case of Metro Cebu, Philippines", *Transportation research interdisciplinary perspectives*, vol. 4, pp. 100078.

McCarthy, L., Delbosc, A., Currie, G. & Molloy, A. 2017, "Factors influencing travel mode choice among families with young children (aged 0-4): a review of the literature", *Transport reviews*, vol. 37, no. 6, pp. 767-781.

McFadden, D. (1972). Conditional logit analysis of qualitative choice behavior.

McFadden, D. (1974) “The measurement of urban travel demand,” *Journal of public economics*, 3(4), pp. 303–328.

McFadden, D. and Train, K. (2000) “Mixed MNL models for discrete response,” *Journal of applied econometrics (Chichester, England)*, 15(5), pp. 447–470.

Misra, S., Li, H. & He, J. 2019, Machine Learning for Subsurface Characterization, 1st edn, Elsevier Science & Technology, San Diego.

Moped (2024) *Wikipedia*. Available at: https://en.wikipedia.org/wiki/Moped (Accessed: 30 January 2024).

Mussone, L. & Changizi, F. 2023, "A study on the factors that influenced the choice of transport mode before, during, and after the first lockdown in Milan, Italy", *Cities*, vol. 136, pp. 104251-104251.

Narkhede, S. (2018) *Understanding AUC - roc curve, Medium*. Available at: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5 (Accessed: 20 February 2024).

Omrani, H. (2015) ‘Predicting travel mode of individuals by machine learning’, *Transportation Research Procedia*, 10, pp. 840–849. doi:10.1016/j.trpro.2015.09.037.

Polamuri (2023) 7 most popular boosting algorithms to improve machine learning model's performance, Dataaspirant. Available at: https://dataaspirant.com/boosting-algorithms/ (Accessed: 12 January 2024).

Raschka, S. (2023) *Stackingcvclassifier: Stacking with cross-validation, StackingCVClassifier: Stacking with cross-validation - mlxtend*. Available at: https://rasbt.github.io/mlxtend/user_guide/classifier/StackingCVClassifier/ (Accessed: 20 February 2024).

Rocca, J. (2019) Ensemble methods: Bagging, boosting and stacking, Medium. Available at: https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205 (Accessed: 11 January 2024).

Shanmugam, L. and Ramasamy, M. (2021) "Study on mode choice using nested logit models in travel towards Chennai metropolitan city," Journal of ambient intelligence and humanized computing, pp. Journal of ambient intelligence and humanized computing, 2021.

Souza, G.F.M.de *et al.* (2021) *Reliability Analysis and Asset Management of Engineering Systems*. 1st edn. San Diego: Elsevier (Advances in Reliability Science).

Stefan Trueck, S.T.R. (2009) Rating Based Modeling of Credit Risk. 1st edn. San Diego: Elsevier Science.

Tamim Kashifi, M. et al. (2022) "Predicting the travel mode choice with interpretable machine learning techniques: A comparative study," Travel, behaviour & society, 29, pp. 279–296.

Trevisan, V. (2022) *Using shap values to explain how your Machine Learning Model Works, Medium*. Available at: https://towardsdatascience.com/using-shap-values-to-explain-how-your-machine-learning-model-works-732b3f40e137 (Accessed: 27 February 2024).

Wang, S., Mo, B. and Zhao, J. (2021) "Theory-based residual neural networks: A synergy of discrete choice models and deep neural networks," *Transportation research. Part B: methodological*, 146, pp. 333–358.

Willis, K.G. (2014) "The Use of Stated Preference Methods to Value Cultural Heritage," *Handbook of the Economics of Art and Culture*, 2, pp. 145–181.

XGBoost (2023) GeeksforGeeks. Available at: https://www.geeksforgeeks.org/xgboost/ (Accessed: 12 January 2024).

Yang, C., Fridgeirsson, E.A., Kors, J.A., Reps, J.M. & Rijnbeek, P.R. 2024, "Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data", Journal of big data, vol. 11, no. 1, pp. 1-17.

Zhang, Z. et al. (2020) "A Customized Deep Neural Network Approach to Investigate Travel Mode Choice with Interpretable Utility Information," Journal of advanced transportation, 2020, pp. 1–11.

Zhao, X. et al. (2020) "Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models," Travel, behaviour & society, 20(C), pp. 22–35.

Zhou, ZH. (2021). Introduction. In: Machine Learning. Springer, Singapore. https://doi.org/10.1007/978-981-15-1967-3_1